Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A proteocistromic atlas of 216 human disease-relevant transcription factors

Gong-Hong Wei

gonghong_wei@fudan.edu.cn

Fudan University Shanghai Cancer Center & MOE Key Laboratory of Metabolism and Molecular Medicine and Department of Biochemistry and Molecular Biology of School Basic Medical Sciences, Shanghai Medi https://orcid.org/0000-0001-6546-9334

Zixian Wang

Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College https://orcid.org/0009-0004-7537-2789

Zenglai Tan

University of Oulu

Xiaonan Liu

Institute of Biotechnology, Molecular Systems Biology, University of Helsinki https://orcid.org/0000-0002-9600-0536

Guowen Duan

Fudan University Shanghai Cancer Center & MOE Key Laboratory of Metabolism and Molecular Medicine and Department of Biochemistry and Molecular Biology of School Basic Medical Sciences, Shanghai Medi

Peng Zhang

Fudan University Shanghai Cancer Center & MOE Key Laboratory of Metabolism and Molecular Medicine and Department of Biochemistry and Molecular Biology of School Basic Medical Sciences, Shanghai Medi

Wenjie Xu

Fudan University Shanghai Cancer Center & MOE Key Laboratory of Metabolism and Molecular Medicine and Department of Biochemistry and Molecular Biology of School Basic Medical Sciences, Shanghai Medi

Binjie Luo

University of Oulu

Longguang Qin

University of Oulu

Yuehong Yang

University of Oulu

Shuangshuang Ma

University of Oulu
Xiayun Yang
University of Oulu
Matias Kinnunen
University of Helsinki
Iftekhar Chowdhury
University of Helsinki
Qin Zhang
University of Oulu
Aki Manninen
University of Oulu
Markku Varjosalo
Institute of Biotechnology, Helsinki Institute of Life Science HiLIFE, P.O Box 56, 00014 University of Helsinki, Helsinki, Finland https://orcid.org/0000-0002-1340-9732

Article

Keywords: Transcription factor, AP-MS/BioID-based proteomics, ChIP-seq, Proteocistromic, Physiological and disease states

Posted Date: March 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8894562/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: There is **NO** Competing Interest.

Abstract

Transcription factors (TFs) execute regulatory programs by integrating signaling inputs with chromatin context, yet their activity has typically been examined either through chromatin occupancy or protein-protein interactions (PPIs), leaving unclear how regulatory information is jointly implemented on the genome. Here, we construct a comprehensive proteocistromic atlas of 216 human TFs within a unified cellular context by integrating affinity-based proteomics with ChIP-seq, mapping over 30,000 high-confidence TF-centered PPIs and showing that TF binding spans 11.66% of the human genome. We develop a quantitative proteocistromic score that integrates DNA-binding activity with effector-domain connectivity, enabling stratification of TFs by functional regulatory potency. High-scoring TFs are enriched for lineage-specifying factors and master regulators of cell fate. These TFs preferentially assemble into cooperative pairs that co-occupy shared regulatory elements, revealing coordinated control modules central to developmental and oncogenic pathways. Within these highly active TF clusters, allele-specific binding events preferentially colocalize with eQTLs and GWAS lead variants, particularly at proximal regulatory elements, directly linking TF binding asymmetry to genetically driven transcriptional varaition. Network-level analyses further delineate synergistic TF pairs and higher-order TF communities that are selectively rewired across cancer cohorts, nominating context-dependent regulatory hubs with biomarker and therapeutic potential. By anchoring genetic variation and transcriptional control within a unified proteocistromic reference, this study defines general organizational principles of TF function on chromatin and establishes proteocistromics as a scalable framework for connecting TF binding, protein interaction networks, and human disease risk.

Introduction

Signal transduction pathways enable cells to sense extracellular and intracellular cues and convert these inputs into coordinated transcriptional responses that govern cell identity, tissue homeostasis, and adaptive or pathological states[1]. At the terminal nodes of these cascades, transcription factors (TFs) function as decisive molecular interpreters, integrating diverse upstream signals to engage chromatin and orchestrate gene expression programs. Dysregulation of TF activity represents a unifying feature of numerous human diseases, including cancer, autoimmune disorders, metabolic syndromes, and neurodegenerative conditions, in which aberrant signaling cascades converge on TF networks to establish maladaptive transcriptional states[2-4]. Recent studies exemplify this principle through signal-responsive TF axes, such as JMJD6-EHF in radioresistant lung cancer and SRF/MRTF-A in hypoglycemia-induced neurodegeneration, that act as critical conduits linking signaling perturbations to disease-driving transcriptional programs[5,6]. Despite their central regulatory roles, TFs have

historically been considered difficult therapeutic targets due to their context-dependent regulation, lack of intrinsic enzymatic activity, and extensive interaction networks[7-9]. Nevertheless, emerging evidence demonstrates that rationally optimized small molecules can directly engage TF activation domains and modulate transcriptional outputs[10]. These advances underscore the necessity of a mechanistic framework that explains how TFs integrate signaling inputs through coordinated chromatin engagement and protein-protein interactions to enable actionable therapeutic intervention.

Historically, TF function has been conceptualized as the integration of two modular properties: DNA-binding domains (DBDs), which recognize sequence-specific genomic motifs, and effector domains (EDs), which recruit cofactors, chromatin remodelers, and signaling components[2-4,11-14]. While more than 1,600 human TFs have been annotated and their DNA-binding specificities extensively cataloged, it is now clear that DNA recognition alone provides an incomplete representation of TF function[15-18]. Instead, transcriptional outputs emerge from combinatorial TF assemblies whose composition, stoichiometry, and activity are dynamically shaped by signaling states, chromatin context, and cellular environment. In this framework, TFs function not merely as genome readers but as regulatory hubs at which kinase cascades, post-translational modifications, chromatin remodeling, and metabolic cues converge[19]. Importantly, disease-associated signaling rewiring can profoundly remodel TF interaction landscapes without altering intrinsic DNA-binding specificity, thereby producing divergent transcriptional outcomes from similar cistromic patterns[20,21]. These observations suggest that a comprehensive understanding of TF function requires a multidimensional perspective that integrates chromatin occupancy with interaction capacity.

However, reconstructing such integrated TF regulatory networks remains a formidable challenge. TFs rarely act in isolation; rather, they operate within multi-protein complexes that incorporate kinases, ubiquitin ligases, chromatin modifiers, and metabolic sensors[22-26]. These TF-centered protein-protein interactions (PPIs) determine whether a given DNA-binding event culminates in transcriptional activation, repression, or functional neutrality. Consequently, identical TF binding profiles can yield distinct gene expression programs depending on the surrounding proteomic environment. Multivalent interactions may further drive the formation of enhancer-associated transcriptional condensates that concentrate TFs, cofactors, and RNA polymerase II, thereby reinforcing context dependence[26-30]. Such regulatory plasticity is particularly evident in cancer and inflammatory diseases, where aberrant signaling rewires TF interaction networks to establish oncogenic or immune-evasive transcriptional states and promote therapeutic resistance. Large-scale initiatives such as the Encyclopedia of DNA Elements (ENCODE) project have generated invaluable maps of TF occupancy and chromatin states across diverse cell types[16,31]. Yet, cistromic data alone do not capture the dynamic PPIs

that enable TFs to interpret upstream signaling inputs. Conversely, proteomic studies have delineated extensive TF interaction networks but often lack direct linkage to genome-wide regulatory activity[20,21,32-36]. Chromatin-bound proteomic strategies, including ChIP-mass spectrometry, partially address this limitation by anchoring interaction measurements to DNA-bound complexes, but these approaches have not been systematically integrated with genome-wide occupancy maps at scale[37]. This methodological disconnect between chromatin-centric and protein-centric analyses has constrained our ability to reconstruct how signal transduction pathways converge at TF nodes to generate disease-specific transcriptional outputs and therapeutic vulnerabilities.

To overcome this limitation, we developed an integrated proteocistromic framework that unifies proteomic and cistromic dimensions of TF regulation. Using the MAC-tag system, we systematically interrogated both stable and transient TF-centered PPIs through affinity purification-mass spectrometry (AP-MS) and proximity-dependent biotin identification (BioID), respectively, while concurrently defining genome-wide TF occupancy by chromatin immunoprecipitation sequencing (ChIP-seq) under tightly controlled expression conditions[38]. To ensure reproducibility and minimize experimental variability, we employed the Flp-In T-REx 293 system to generate isogenic, stable, and inducible cell lines with precisely regulated TF expression[36]. Although prior studies have demonstrated the utility of this system for reconstructing large-scale TF interactomes, its potential for integrative transcriptional network analysis has remained largely unexplored[35,39]. Importantly, AP-MS captures stable transcriptional complexes, whereas BioID detects transient or spatially restricted interactions that frequently mediate signal-dependent regulation[38], thereby providing complementary views of TF-centered regulatory architecture. Applying this strategy to 216 disease-relevant human TFs, we generated a comprehensive proteocistromic atlas comprising more than 30,000 high-confidence interactions alongside genome-wide TF binding spanning approximately 12% of the human genome. This resource enables systematic stratification of TFs according to interaction capacity, chromatin engagement, cooperativity, and sensitivity to genetic variation. Moreover, it reveals how distinct proteocistromic states correspond to regulatory dominance, disease association, and cancer-specific TF communities. By establishing proteocistromics as both a conceptual and analytical framework, our study advances a systems-level understanding of TF-mediated signal transduction and provides a foundation for identifying transcriptional dependencies and therapeutically actionable targets in human disease.

Results

Proteocistromic landscape of transcription factors in human cells

To systematically delineate how TFs integrate protein interaction networks with chromatin engagement, we performed a comprehensive proteocistromic analysis of 216 human TFs. This framework combined genome-wide DNA-binding profiles generated by ChIP-seq with TF-centered PPI networks mapped using AP-MS and BioID (**Figure S1a; Table S1**).

Among the profiled TFs, 157 (73%) were expressed above a stringent threshold of 1 transcript per million (TPM) in Flp-In™ T-REx 293 cells (**Figure 1a**), ensuring robust and interpretable downstream measurements. These TFs represent 31 recognized DBD families, together with six TFs whose DBDs remain unclassified (**Figure 1b**). Notably, 94% of the analyzed TFs have been implicated in human diseases[40] (**Figure 1c**), underscoring the broad biomedical relevance of this resource and its potential to illuminate molecular mechanisms underlying diverse pathologies. Within the resulting interaction network, ARID3A and KMT2D emerged as highly connected hubs, each interacting with more than 100 TFs (**Figure 1d**), consistent with their established roles as transcriptional coregulators[41] and global regulators of gene expression programs.[42]

To enable systematic and comparable interactome profiling, we generated inducible MAC-tagged TF constructs in Flp-In™ T-REx 293 cells, facilitating rapid establishment of cell lines with uniform expression levels[35,38,43] (**Figure 1e**). Integration of AP-MS and BioID datasets yielded a comprehensive TF interaction landscape comprising 32,931 high-confidence interactions (HCIs) involving 4,156 unique proteins (**Table S2**). On average, each TF engaged in 133 interactions detected by BioID and 36 interactions identified through AP-MS (**Figure 1f**), reflecting the complementary detection principles of proximity labeling and affinity purification. Gene Ontology enrichment analysis demonstrated strong overrepresentation of transcriptional regulation, RNA/DNA binding, and signal transduction processes among TF-associated proteins (**Figure 1g**). Importantly, both the scale and coverage of this network substantially exceed those of prior yeast two-hybrid[17] and AP-MS-based studies.[32] Moreover, 63% of interactions were not previously annotated in public protein interaction databases (**Table S2**), highlighting the substantial expansion of TF-associated interaction space achieved here.

Despite their complementary methodologies, only 3.9% (1,285 of 32,931) of interactions were detected by both AP-MS and BioID, representing greater concordance than the 2.49% overlap reported in our earlier dataset (200 of 8,039 interactions).[35] Strikingly, we identified more than 6,600 TF-TF interactions, with each TF interacting, on average, with over 31 other TFs, emphasizing the extensive combinatorial connectivity of transcriptional networks. Of these TF-TF interactions, 95.5% were detected by BioID, whereas 9.8% were captured by AP-MS and 5.4% by both approaches. This strong enrichment in proximity-based detection reflects the transient and low-affinity nature of many TF-TF associations

and underscores the critical value of BioID for mapping dynamic regulatory assemblies.[32,44]

To assess robustness and cross-context generalizability, we performed independent AP-MS validation in the prostate cancer cell line 22Rv1. Ten TFs spanning five DBD families (TEAD, ETS, RFX, NFI, and HOXA; two TFs per family) were randomly selected for interactome profiling. AP-MS analysis in 22Rv1 cells identified 250 total interactions, of which 107 (42.8%) overlapped with interactions detected in HEK 293 cells (**Figure S1b**; **Table S2**). This substantial concordance supports both the technical reproducibility of our workflow and the biological relevance of these interactions across distinct cellular contexts.

Having established a large-scale and reproducible TF interactome, we next characterized the corresponding cistromic landscape. ChIP-seq profiling revealed that TF binding sites were distributed across all human chromosomes except chromosome Y. Peak calling and motif analyses confirmed high data quality, with the number of binding sites per TF ranging from 43 (ZBTB33) to 131,094 (ZIC3), with a median of 26,154 sites. Collectively, TF binding sites encompassed 360.96 million base pairs, corresponding to 11.66% of the human genome. Among DBD families, C2H2 zinc finger, Homeodomain, and ETS families exhibited the greatest aggregate genomic coverage, each contributing more than one million peaks (**Figure 1h**).

Motif analysis revealed both canonical and noncanonical (secondary) motifs for many TFs, consistent with widespread co-binding and tethered-binding mechanisms in transcriptional regulation.[16,45,46] Across 206 TFs, we identified 261 canonical and 654 noncanonical motifs, corresponding to 255 putative co-binding and 399 tethered-binding interaction types (**Table S3**). These data indicate that tethered binding is more prevalent than direct co-binding, consistent with previous observations.[45] Importantly, this *in vivo* binding resource substantially complements prior large-scale *in vitro* TF-DNA interaction studies[47-49] by providing chromatin-contextualized specificity information.

We then integrated the proteomic and cistromic layers to directly connect physical interactions with chromatin co-occupancy. This integrative analysis identified 15 TF pairs supported by both ChIP-seq co-binding and PPI evidence, including nine co-binding pairs (e.g., ELF1-SP1[50] and NFYC-SP2[51]) and five tethered-binding pairs detected by BioID. Notably, the MYC-MAZ interaction[52] was classified as tethered binding by both AP-MS and BioID (**Figure 1i**). Among the 654 inferred TF dimer pairs, MYC-MAX, GATA2-SMAD4, and ELF1-SP1 are curated in the IntAct database[53], and recovery of MYC-MAX and NFIC-SP1 interactions further substantiates the technical robustness and biological relevance of our network. Beyond identifying cooperative TF pairs, motif-based clustering refined

subclasses within several large TF families previously defined by primary motif preferences[54] (**Figure S1c**), suggesting additional layers of regulatory specialization.

Collectively, these integrative analyses establish an unprecedented proteocistromic landscape of human TFs, uncovering extensive physical connectivity and diverse modes of regulatory engagement. This resource provides a comprehensive framework for dissecting TF cooperativity and mechanistic specificity in human cells and more broadly, lays the foundation for future investigations into context-dependent transcriptional regulation in both physiological and pathological settings.

Proteomic and cistromic coordination of transcription factors

The regulatory behavior of TFs is governed by the coordinated functions of their EDs and DBDs, reflecting their dual roles in signal integration and chromatin engagement. Upon binding to regulatory elements, TFs modulate gene expression through direct cooperation with other TFs and through assembly of higher-order transcriptional complexes that recruit cofactors to chromatin. These multilayered interactions finely tune transcriptional outputs and confer regulatory plasticity across cellular contexts, underscoring the architectural complexity of TF-cofactor networks.[2,4,55-59] Despite the identification of thousands of coregulators, systematic delineation of cofactors recruited by individual TFs remains challenging, largely because these interactions are transient, dynamic, and highly context dependent.[2,16,33]

To overcome this limitation, recent large-scale efforts have quantitatively profiled ED activity across diverse TFs.[31,60] Building on these datasets, we integrated measurements from the Bintu group[31,60] to classify TFs according to ED function as activation domains (ADs), repression domains (RDs), or bifunctional domains (Bif). In parallel, we annotated interacting prey proteins as activators (A), repressors (R), or dual-function (AR) cofactors based on Uniprot annotations,[61] and constructed an integrated TF-TF and TF-cofactor interaction network (**Figures 2a and S2a**).

To extract higher-order structure from this network, we derived quantitative PPI feature matrices and applied k-means clustering to stratify TFs into three groups, namely P1 (high), P2 (medium), and P3 (low), representing graded levels of ED-associated interaction capacity (**Figure 2b**). Although silhouette analysis indicated that $n = 2$ was statistically optimal, we selected $n = 3$ to better capture the biological continuum of TF interaction states. Importantly, the intermediate P2 cluster comprised TFs with moderate yet functionally meaningful interaction profiles, supporting the biological relevance of a three-tiered stratification (**Figures S3a-S3d**).

Because interaction capacity must ultimately be manifested through chromatin engagement, we next asked whether these interaction-defined TF groups exhibit distinct cistromic architectures. TF binding across promoters, enhancers, and insulator elements constitutes the structural foundation of gene regulatory networks and often reflects intrinsic regulatory properties of TFs.[62,63] Unsupervised hierarchical clustering of TF occupancy across these genomic elements identified five reproducible cistromic clusters: C1, enriched at promoters; C2, preferentially associated with enhancers; C5, dominated by CTCF-proximal binding; C3, displaying relatively balanced occupancy across all three elements; and C4, exhibiting mixed binding patterns (**Figure 2c-2e**). Cross-tissue analyses further demonstrated that several clusters were broadly active across both normal and malignant tissues. For example, C1 TFs displayed elevated activity in urinary and digestive systems under both physiological and cancer conditions (**Figures 2f and 2g**), indicating that cistromic architecture is partially conserved across tissue states.

To assess the robustness and generalizability of these cistromic classifications, we repeated the clustering analysis using independent HEK293 ChIP-seq datasets, including ENCODE resources.[64] Genome-wide TF binding again segregated into five analogous clusters (C1–C5) (**Figure S2b**), supporting the reproducibility of this architectural stratification. In our dataset, 64% of TFs (131/206) were assigned to C1 and C3, comparable to the distribution in public datasets (83%, 227/272). Notably, 3 of 11 TFs in our C5 cluster belonged to the C2H2 zinc finger (ZF) family, whereas this proportion reached 94% (16/17) in public datasets, consistent with the enrichment of C2H2 ZF proteins at CTCF-associated regions and with CTCF itself being a C2H2 ZF TF. Moreover, among the 18 TFs shared between our dataset and public HEK293 data, more than half (10/18) were assigned to the same cistromic cluster, with the majority (8/10) residing in C1 (**Figure S2c**), further underscoring the reproducibility of our classification framework.

Having established concordant interaction-based and cistromic stratifications, we next examined whether interaction capacity quantitatively associates with chromatin-binding strength. TFs with higher frequencies of TF-TF and TF-cofactor interactions exhibited marked enrichment of high-confidence chromatin binding events, with approximately 5- to 10-fold increases, particularly at transcription start sites (TSSs) (**Figures 2h, 2i and S2d-S2f**). These findings indicate a tight functional coupling between ED-mediated interaction capacity and DBD-mediated regulatory signal intensity. Furthermore, TFs preferentially associated with distal regulatory elements displayed greater interaction propensities than those primarily engaged at proximal promoters (**Figure 2j**), suggesting that long-range regulatory programs require more extensive cofactor recruitment and cooperative assembly.

Collectively, by integrating quantitative proteomic interaction profiles with genome-wide cistromic landscapes, our analysis reveals coordinated coupling between TF effector-

domain interaction capacity and chromatin-binding architecture. This multi-omic framework resolves distinct TF functional states and provides systems-level insight into how TFs orchestrate gene regulatory networks through tightly linked protein-protein interaction and genomic engagement.

Proteocistromic landscapes define hierarchical transcription factor clusters

Within the global TF-cofactor interaction network, members of the transcriptional corepressor family, particularly transducing-like enhancer of split (TLE) proteins, emerged as dominant partners of RD TFs. Specifically, 50 of 99 RD TFs engaged TLE proteins, accounting for more than 59% (145/244) of all TF-TLE interactions detected (**Figures 3a and 3b**). This pronounced enrichment indicates that TLE recruitment constitutes a central architectural feature of RD-mediated transcriptional repression programs and suggests that discrete corepressor modules may define specific proteocistromic TF states.

Consistent with this interaction bias, ChIP-seq analyses demonstrated robust chromatin co-localization between TLE proteins and TLE-interacting RD TFs. In contrast, RD TFs lacking detectable TLE interactions exhibited substantially reduced genomic overlap (**Figure 3c**). These results establish a strong concordance between physical protein-protein interactions and coordinated genomic occupancy, underscoring the functional coherence of the proteomic network at the chromatin level.

Notably, although the effector-domain activity of NFIB remains incompletely defined, NFIB participates in PPIs with other NFI family members harboring well-characterized repression domains, supporting functional conservation within the NFI family[65] (**Figure 3d**). Notably, NFIA, NFIB, TLE1, and TLE3, each of which engages in reciprocal interactions, displayed coordinated chromatin occupancy at NFI-occupied loci (**Figures 3e and S4a**). These findings reveal a tightly integrated regulatory module in which proteomic connectivity and cistromic co-occupancy reinforce lineage-specific transcriptional programs.

Collectively, these observations suggested that TF functional states cannot be fully resolved by interaction profiles or chromatin binding patterns alone. We therefore hypothesized that simultaneous interrogation of proteomic and cistromic dimensions would enable more refined functional stratification of TFs. To implement this framework, we constructed a comprehensive set of proteocistromic features by integrating quantitative interaction metrics including both TF-cofactor interactions (TCIs) and TF-TF interactions (TTIs)—with measures of chromatin occupancy and regulatory intensity. Specifically, we assessed the distribution and enrichment of high-confidence binding events, defined as ≥10-fold ChIP-seq peak enrichment, across promoters, enhancers, and CTCF-associated regions (**Figure 3f**).

Unsupervised clustering based on these integrated features (see **Methods**) resolved TFs into three distinct functional clusters (**Figures 3g and S3e-S3j**). Cluster a was characterized by strong contributions from TCIs, TTIs, and DBD activity, identifying TFs with dense interaction networks and broad chromatin engagement. In contrast, clusters b and c displayed progressively reduced interaction densities and diminished regulatory footprints, consistent with more restricted functional influence. Notably, ETS family TFs were highly enriched in cluster a, underscoring their central roles in transcriptional regulation (**Figure 3h**).

Functional enrichment analysis further revealed pronounced divergence among clusters. TFs in cluster a were significantly associated with pathways governing cell fate determination, DNA-templated transcription, transcriptional dysregulation in cancer, and glandular morphogenesis, whereas clusters b and c showed comparatively weaker enrichment across these core regulatory programs (**Figures 3i and S4b**).

Taken together, these findings position cluster a TFs as dominant organizers of gene regulatory networks and demonstrate that proteocistromic integration provides a powerful framework for uncovering hierarchical functional organization within the transcriptional regulatory landscape.

Characterizing transcription factor pairs and multilayer cooperativity

Having established a proteocistromic framework for functional TF stratification, we next investigated how cooperative interactions among TFs further shape regulatory output. Cooperative TF binding represents a fundamental mechanism through which combinatorial signaling inputs are integrated at cis-regulatory elements to generate precise, context-dependent transcriptional responses[58,66]. Importantly, such cooperativity can arise through multiple, partially overlapping molecular mechanisms[67].

First, direct protein-mediated TF-TF interactions enable the formation of dimers or higher-order complexes that bind DNA cooperatively, frequently recognizing palindromic or repeated motifs, as exemplified by the BMAL-CLOCK heterodimer.[68] Second, DNA-facilitated cooperativity occurs when TFs with limited intrinsic affinity in solution assemble into stable complexes upon engagingadjacent DNA motifs, as observed for SOX2 and OCT1.[69] Third, DNA-mediated allosteric interactions arise when binding of one TF induces local alterations in DNA shape or flexibility that enhance recruitment of a partner factor, as demonstrated for NKX2.5 and TBX5.[66] Finally, nucleosome-mediated cooperativity reflects indirect coupling of TF occupancy through chromatin remodeling or competitive nucleosome displacement, whereby binding of one factor increases local accessibility for another, as reported for Gal4 and LexA.[70]

Although high-throughput *in vitro* binding assays such as CAP-SELEX (Systematic Evolution of Ligands by Exponential Enrichment coupled with Consecutive Affinity Purification) have enabled systematic identification of TF pairs recognizing composite binding sites,[69] the prevalence and functional relevance of TF cooperativity in vivo remain incompletely defined. To address this gap, we performed a comprehensive in vivo assessment of TF cooperativity and evaluated its biological relevance in tumor contexts.

Using an established computational pipelineto to analyze spatial relationships among TF binding events within the three previously defined TF clusters,[71] we identified 184 significant TF pairs, comprising 159 relaxed-spacing and 25 constrained-spacing pairs (**Table S4**). Notably, 68% of TFs participating in these interactions belonged to cluster a. Consistent with this enrichment, ranking TFs by proteocistromic scores revealed that, irrespective of interaction type, cooperative TFs were preferentially concentrated among the most active subset within cluster a (**Figure 4a**). These results indicate that cooperative binding is strongly associated with transcriptionally dominant regulatory programs.

Given this strong association with regulatory dominance, we next asked whether these TF pairs contribute to tumor-related phenotypes. MAX and FOXJ2 emerged as the top-ranked TFs in relaxed- and constrained-spacing pairs, respectively. Among MAX-associated interactions, members of the KLF family, particularly KLF4, KLF5, and KLF8 were highly represented, with KLF5 exhibiting the highest proteocistromic score. FOXJ2 also showed frequent pairing with KLF4 (**Figures 4a-4c**).

To determine clinical relevance, we examined survival associations in both the TCGA KIRC (kidney renal clear cell carcinoma) cohort and our independent RCC cohort. Patients with high co-expression of MAX and KLF5 exhibited significantly improved overall survival compared with those with low co-expression (**Figures 4d, S5a and S5b**). A similar survival advantage was observed for patients with high FOXJ2-KLF4 co-expression (**Figures 4e, S5a and S5C**). Functional validation in three RCC cell lines (786-O, ACHN, and Caki-1), demonstrated that co-overexpression of MAX-KLF5 or FOXJ2-KLF4 significantly suppressed cell migration and reduced 3D colony formation (**Figures 4f-4g and S5d**). Together, these findings support cooperative tumor-suppressive roles for these TF pairs.

To further elucidate the molecular basis of TF cooperativity, we integrated proteomic interaction data with TF-pair predictions. Among the 184 candidate pairs, 23 were detected in proteomic assays. Of these, six pairs primarily from the NFI and RFX families, were identified by both AP-MS and BioID; TEAD1-TEAD4 was uniquely detected by AP-MS; and the remaining 16 pairs were detected exclusively by BioID (**Table S4**). Comparison with CAP-SELEX datasets revealed six overlapping TF pairs, all identified by BioID (**Figure S6a; Table S4**). This pattern is consistent with the notion that many TF-TF interactions are

transient rather than stable.[35] Because BioID captures proteins within an approximately 10-nm radius,[72,73] these TF pairs may not necessarily form direct physical contacts but instead colocalize within shared regulatory regions. Such spatial proximity could facilitate functional cooperation through DNA bending or higher-order chromatin architecture, enabling coordinated transcriptional regulation. Collectively, these findings underscore the multilayered and dynamic nature of TF cooperativity within regulatory genomic landscapes in vivo.

Consistent with this interpretation, four TF pairs were jointly identified in both ChIP-seq and CAP-SELEX datasets, all belonging to the ETS family and characterized by relaxed spacing constraints. Motif enrichment analysis of the top 500 ChIP-seq peaks revealed corresponding heterodimeric ETS motifs (**Figure S6b**), highlighting family-specific cooperative binding architectures. However, because CAP-SELEX is performed *in vitro* and may not fullly recapitulate chromatin-dependent interactions *in vivo*,we further compared our proteomic interaction dataset with previously reported TF-TF interactions not detected by CAP-SELEX,[69] identifying 104 overlapping interactions. Notably, several well-established TF partnerships, including E2F3-SP1,[74] ELK1-SRF,[75] ERG-FLI1,[76] FOS-JUN,[77] FOXA1-GATA3,[78] and MYC-MAX,[52] were absent from CAP-SELEX datasets, underscoring the limitations of in vitro binding assays in capturing context-dependent TF cooperation.

Finally, systematic evaluation of these 104 TF pairs using TCGA Pan-Cancer transcriptomic data revealed that the majority exhibited significant positive co-expression across multiple cancer types (**Figure S5e**), indicating that these interactions are not restricted to specific tumor contexts. Beyond the well-characterized ERG-FLI1 and FOS-JUN pairs, additional combinations including FOXJ2-ELF1, POU2F1-NR2C2, POU2F1-SP1, and SPI1-FLI1, also demonstrated robust positive correlations across diverse tumor types (**Figure 4h**). Collectively, these results demonstrate that our in vivo-derived TF interaction landscape complements *in vitro* studies and reveals the pervasive involvement of TF cooperativity in oncogenic regulatory programs.

Characterizing shared binding sites and interacting transcription factors

While the analyses above focus on discrete TF pairs, transcriptional regulation in vivo is rarely governed by isolated binary interactions. Instead, it is typically orchestrated by higher-order assemblies of multiple TFs that form dense regulatory clusters, frequently positioned near cohesin anchor sites in human cells.[79] Accordingly, gene expression is more commonly driven by coordinated multi-factor occupancy than by pairwise interactions alone.[80] Building on this concept, we next asked whether multiple TFs co-occupy the same genomic loci and simultaneously engage in direct or indirect protein-

protein interactions (**Figure 5a**), thereby forming integrated regulatory modules that couple DNA binding with proteomic connectivity.

To address this question, we integrated TF binding profiles with TF-TF interaction data and identified 72,964 genomic regions bound by multiple TFs that also exhibit TF-TF interactions. We define these loci as SBSI-TF regions (Sharing Binding Sites and Interacting TFs). Compared with non-SBSI-TF regions, SBSI-TF regions were significantly enriched at promoters (**Figure 5b**), indicating that coordinated TF assemblies preferentially localize to transcription initiation sites.

Given this pronounced promoter enrichment, we next examined the properties of promoters overlapping SBSI-TF regions. We identified 16,739 promoters overlapping SBSI-TF regions, compared with 32,049 promoters associated with non-SBSI-TF regions. The SBSI-TF-associated promoters corresponded to 10,071 genes. Notably, housekeeping genes comprised only a small fraction of these genes (12.1%, 1,219/10,071), suggesting that SBSI-TF regions are not predominantly associated with constitutively expressed loci.

Integration of GTEx transcriptomic data further refined this distinction. While housekeeping genes are broadly expressed across tissues, most SBSI-TF-associated genes clustered within dashed-line box 1 (**Figure 5c**), characterized by either consistently high or consistently low expression with limited tissue specificity. In contrast, housekeeping genes located in dashed-line boxes 2 and 3 were predominantly highly expressed in brain tissues. Thus, SBSI-TF-associated promoters appear to represent regulatory programs that are stable but not uniformly ubiquitous across tissues. In addition, promoters within SBSI-TF regions exhibited a markedly higher prevalence of CpG islands compared with non-SBSI-TF promoters (Fisher's exact test, $p < 2.2 \times 10^{-16}$; **Figure 5d**), indicating a CpG-rich promoter architecture often associated with transcriptional competence and regulatory plasticity.

Because dense TF occupancy is a defining feature of highly active regulatory elements, we next examined the relationship between SBSI-TF regions and previously defined high-occupancy target (HOT) regions, which are characterized by exceptionally dense TF binding.[81] We quantified overlaps with HOT regions, promoters, and CpG islands by both the number and proportion of overlapping bases, defining overlap as any shared nucleotide between regions. Although a subset of SBSI-TF regions overlapped annotated HOT regions (5,473 of 72,964),[81] this represented only a minority of all SBSI sites. Nevertheless, compared with non-SBSI regions, SBSI-TF regions showed significantly greater overlap with HOT regions and were more frequently co-localized with promoters and CpG islands (**Figure 5e**).

These results indicate that SBSI-TF regions constitute a complementary class of highly active regulatory elements that are distinct from, yet partially overlapping with, canonical HOT regions. Importantly, the identification of this expanded regulatory landscape is likely enabled by the scale of the TF binding datasets and, critically, by the integration of protein–protein interaction information. Consistent with this interpretation, TFs belonging to proteocistromic cluster a displayed dominant occupancy within SBSI-TF regions (**Figure 5f**), further supporting their cooperative and regulatory prominence.

While the analyses above establish the genomic and regulatory features of SBSI-TF regions, cooperative TF binding ultimately depends on molecular compatibility among interacting proteins. Because protein interactions are frequently constrained by structural complementarity, we next investigated whether specific domain architectures are associated with TF-TF interactions.

Proteins harboring particular combinations of structural domains are known to interact preferentially, as reflected by extensive domain-domain associations in human interactomes.[32,34] Accordingly, we mapped Pfam domains[82] across the TF interactome and identified 951 significant domain associations (FDR < 0.01) among 13,120 TF-protein interactions (**Figure S7a; Table S5**). Although these associations do not necessarily imply direct physical contact between domains, they reveal recurrent structural configurations that may predispose TFs to cooperative engagement.

Notably, the two most prevalent TF DBD families, homeodomain and C2H2 zinc fingers, exhibited strong intra-family associations (FDR = 0 and FDR = 1.78×10^{-99}, respectively) (**Figures S7a and S7b**), suggesting preferential self-family interaction patterns. In addition, highly significant cross-family associations were observed, including interactions between ETS and homeodomain TFs (FDR = 0; **Figure S7c**), indicating that cooperative assemblies can arise from complementary structural frameworks.

Collectively, these findings reveal that SBSI-TF regions are not only defined by shared genomic occupancy and interaction connectivity, but are also supported by underlying structural compatibilities among TF domain architectures. Together, this integrative view links chromatin co-occupancy, protein interaction networks, and domain-level organization to explain how higher-order TF assemblies are stabilized at shared regulatory loci.

Proteocistromic profiles shape genetic variation in transcription factor chromatin binding

While domain-mediated interactions provide a structural framework for TF cooperativity, naturally occurring genetic variation offers a complementary in vivo strategy to interrogate functional TF binding. ChIP-seq enables high-resolution mapping of TF occupancy across the genome, and non-reference read alignments frequently harbor single-nucleotide mismatches arising from heterozygous variants. These mismatches can

produce allelic imbalances in read counts, thereby enabling the identification of allele-specific binding (ASB) events.[83] Leveraging this principle, we systematically quantified ASBs across three proteocistromic TF clusters to evaluate how integrated TF features, defined by DBDs and Eds, modulate genetic sensitivity at single-nucleotide resolution. Across TFs with defined proteocistromic features, ASB burden varied widely, ranging from 0 to 608 events, with a median of 104.5 (**Table S6**), indicating substantial heterogeneity in genetically modulated chromatin binding.

Importantly, this variability could not be explained solely by DBD classification. Despite comparable base-substitution frequencies, TFs within the same DBD family exhibited markedly divergent ASB burdens (**Figure 6a**), demonstrating that DBD identity alone does not adequately predict functional variability in chromatin engagement. Instead, these observations support a proteocistromic model in which cooperative activity emerges from the integrated contributions of DBDs and ED-mediated interactions. Consistent with this framework, ASB sites bound by multiple TFs were significantly enriched for overlap with expression quantitative trait loci (eQTLs) across diverse tumor types in TCGA and normal tissues in GTEx (**Figures 6b and 6c**). This enrichment suggests that variants embedded within cooperative TF assemblies are more likely to exert measurable regulatory effects on gene expression.

To further delineate the regulatory landscape of TF-associated ASBs, we integrated Manhattan plotting with co-localization analyses between ASBs and cis-eQTLs in TCGA and GTEx. This genome-wide approach identified rs655530 and rs117990130 as prominent candidate regulatory variants. Specifically, rs655530 was associated with binding by 40 TFs and functioned as a significant eQTL across 19 tumor types in TCGA, whereas rs117990130 was linked to 47 TFs and acted as a significant eQTL across 47 tissues in GTEx (**Figures 6d–6f**). These loci thus represent convergent hotspots where cooperative TF binding intersects with genetically mediated transcriptional regulation.

We next sought to uncover the mechanistic basis underlying these associations. Motif screening using the enhancer element locator (EEL) algorithm[84] predicted allele-dependent motif disruptions for FOXC1, NR2E1, and ZNF75A at rs655530. Although direct ChIP-seq evidence of binding by these TFs at this locus was not observed, PPI analysis revealed that FOXC1 interacts with 11 additional TFs, raising the possibility that it contributes to indirect or cooperative recruitment within higher-order TF complexes at this site (**Figures 6g, 6i, and 6k**). Similarly, rs117990130, ranked as the top GTEx eQTL, displayed motif-predicted ASBs for five TFs (GMEB2, ID4, MESP1, NHLH1, and TCF4). PPI data further indicated that GMEB2 directly interacts with six TFs (CPHXL, ELK3, KLF3, KLF12, KLF15, and RFX3), suggesting that GMEB2-centered interaction networks may facilitate coordinated transcriptional regulation at this locus (**Figures 6h, 6j and 6l**).

Collectively, these analyses underscore the importance of network context in determining the functional consequences of ASB variants.

Having established locus-specific mechanistic examples, we next examined cluster-level trends. TFs within the highly active cluster a consistently exhibited a greater ASB burden (**Figure 6m**), suggesting that ASB frequency may serve as a quantitative proxy for regulatory potency. This interpretation is reinforced by the frequent co-localization of ASBs with GWAS lead SNPs and eQTLs (**Figure 6m**), indicating that these genomic regions are particularly sensitive to genetically modulated transcriptional control. Moreover, ASBs were predominantly enriched within proximal promoter regions (**Figure 6n**), further supporting their direct relevance to transcriptional initiation. Among all TFs analyzed, PAX9 displayed the highest ASB burden (n = 608), with 65% overlapping GTEx eQTLs, 27% overlapping TCGA eQTLs, and 10% coinciding with GWAS lead variants (**Figure S8a-S8c**), highlighting its prominent and genetically sensitive regulatory role.

Finally, we experimentally validated these genome-wide patterns using IP-SNP-seq in HEK293 cells. ASBs specific to cluster a exhibited significantly stronger allelic imbalance than those in cluster b/c or those shared across clusters, with the largest effects observed for cluster a-restricted ASBs, as quantified by absolute binding allelic bias (|BAB|) and Cliff's delta (**Figures 6o and 6p**). Paired IP-SNP-seq analyses in HEK293 and 22Rv1 cells further demonstrated that allelic imbalance at cluster a sites was largely preserved across cell types, whereas cluster b/c-associated, shared, and PAX9-related ASBs showed modest but significant increases in |BAB| in prostate cancer 22Rv1 cells (**Figure S8d and S8e**). These findings indicate that motif-restricted ASBs, particularly those linked to cluster a TFs and PAX9, not only reflect intrinsic TF binding preferences but also capture context-dependent amplification of TF activity in malignant cellular environments.

Together, these data establish ASBs as sensitive genetic readouts of proteocistromic TF network activity, linking single-nucleotide variation to cooperative chromatin binding and context-specific transcriptional regulation.

Community-cancer interactions

Building upon the genetic evidence supporting coordinated TF activity, we next sought to determine whether proteocistromic TFs assemble into higher-order interaction communities that are preferentially engaged in cancer. In malignant contexts, TFs are frequently rewired into context-specific interaction modules that couple aberrant signaling pathways to oncogenic transcriptional programs.[8,85] Rather than acting in isolation, cancer-associated TFs tend to interact selectively and cooperatively, forming tumor-context–specific regulatory communities. Functionally related proteins often organize into densely interconnected clusters that represent multiprotein complexes or coordinated

signaling pathways essential for tumor initiation and progression.[34,86] Despite this conceptual framework, the higher-order organization of cancer-associated TF interactions, and the extent to which genetic perturbations reshape community architecture and complex assembly, remains incompletely understood, largely due to limited resolution of the underlying network structure. Moreover, TFs and their interactors frequently form tightly connected mudules that operate as regulatory units critical for oncogenesis and disease progression.[87-89] Elucidating the structure and functional composition of TF-centered interaction communities in cancer may therefore uncover novel mechanistic insights and therapeutic vulnerabilities.

To systematically characterize these communities, we interrogated TF interaction structures within the global PPI network and evaluated their associations with cancer (**Figure 7a; Figure S9**). Using unsupervised Markov clustering (MCL),[90] we analyzed PPI networks encompassing 212 TFs and identified 81 distinct communities connected by 734 inter-community interactions. We next assessed the enrichment of cancer-relevant genes within these communities. Integration with the CancerMine database[91] revealed that 46.7% of the 1,941 genes represented in the TF-centered PPI network have documented associations with cancer (**Figure 7b**). Across all communities, we identified 1,730 cancer-associated community-community interactions spanning 425 distinct cancer types, which were subsequently consolidated into 10 major human cancer systems. Notably, both breast cancer and hepatocellular carcinoma engaged more than 45 distinct TF-centered communities (**Table S7**), indicating extensive network-level involvement in these malignancies. Together, these findings underscore the power of community-level analyses to capture cooperative network behaviors that are not apparent from single-gene perspectives, particularly in complex, heterogeneous diseases such as cancer.[34]

Given this widespread community engagement, we next examined how the three previously defined clusters comprising 184 proteocistromic TFs were distributed across cancer-associated communities. Communities linked to a broader spectrum of cancer types were significantly enriched for proteocistromic TFs (**Figures 7c and 7d; Table S7**), highlighting their pervasive involvement in oncogenic processes and suggesting that they occupy central positions within cancer-specific TF interaction networks.

To further pinpoint key regulatory drivers, we prioritized TFs embedded within highly interconnected cancer-associated clusters. Among all community–community interactions, 28 TFs emerged as central nodes within densely connected cancer-associated modules, consistent with potential roles as regulatory hubs in tumor-related transcriptional programs (**Figures 7e and 7f**). Pathway enrichment analysis demonstrated that these hub TFs are predominantly involved in cell fate determination and transcriptional

dysregulation in cancer, reflecting their broad and multifaceted contributions to oncogenic signaling pathways (**Figure 7g**).

Finally, we investigated whether these hub TFs exhibit cancer-type-specific expression patterns. Using TCGA transcriptomic data and Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction, we found that expression of the 28 hub TFs alone was insufficient to discriminate among cancer types. In contrast, TFs outside this core hub group displayed more cancer-type-restricted expression patterns and enabled clearer separation of tumor entities (**Figure 7h**). These findings suggest that the identified hub TFs participate in fundamental regulatory programs shared across diverse malignancies, consistent with roles in core oncogenic processes. The contrasting expression behaviors between hub and non-hub TF subsets further highlight the potential of combining broadly active regulatory hubs with context-specific TFs to construct complementary biomarker panels for cancer classification and to inform precision therapeutic strategies.

Collectively, our analyses demonstrate that TF-centered interaction communities constitute an organized and cancer-relevant network architecture. Dissecting these higher-order regulatory communities provides critical insight into how coordinated transcriptional programs drive cancer heterogeneity and progression, and reveals new opportunities for network-informed therapeutic intervention.

Discussion

Recent advances in transcription factor (TF) biology have reshaped the classical view of TFs as static DNA-binding regulators, positioning them instead as dynamic integrators of signaling inputs, chromatin landscapes, and protein interaction networks.[92] This conceptual shift reframes TFs from “undruggable” DNA-binding entities to modular regulatory nodes whose function can be modulated through effector-domain interactions, cooperative assembly, and chromatin engagement.[8,93] A central challenge in contemporary transcription biology is therefore not merely defining TF binding specificity, but understanding how TFs integrate signaling dynamics, interaction capacity, chromatin context, and genetic variation into coherent regulatory states.

This evolution builds upon systematic dissection of intrinsic DNA-binding specificity and motif syntax,[47,48,94] which revealed that TFs frequently recognize multiple sequence configurations and exhibit context-dependent binding preferences.[94] Large-scale surveys of human TF variation further demonstrated that coding variants commonly alter DNA-binding affinity and specificity, expanding the regulatory landscape beyond canonical motifs.[95] Collectively, these findings collectively establish that intrinsic sequence recognition is necessary but insufficient to explain TF function in vivo.

To address this gap, we developed a proteocistromic framework that systematically integrates TF-centered interaction networks with genome-wide chromatin occupancy under tightly controlled experimental conditions. Leveraging an inducible Flp-In T-REx system[105-107] in combination with AP-MS and proximity labeling (BioID), we captured complementary interaction modalities, including stable assemblies identified by AP-MS and transient or spatially proximal associations detected by BioID. These interaction profiles were further integrated with ChIP-seq to delineate chromatin engagement. Short-term induction minimized adaptive transcriptional rewiring, enabling a direct and temporally resolved linkage between interaction capacity and chromatin occupancy. This integrative strategy extends prior quantitative proteomic analyses demonstrating dynamic remodeling of TF complexes during differentiation[96] and context-dependent pioneer factor occupancy governed by protein-intrinsic and chromatin features[97]. Importantly, by experimentally unifying interaction and occupancy layers, our framework resolves tethered-binding scenarios in which chromatin recruitment is mediated by interaction partners rather than direct motif recognition.

Quantitative integration revealed a robust coupling between effector-domain interaction capacity and cistromic architecture. TFs with dense TF-TF and TF-cofactor interaction networks exhibited stronger chromatin occupancy and preferential distal enhancer engagement, consistent with the architectural demands of enhancer-driven transcription. In contrast, TFs with limited interaction capacity displayed more restricted binding profiles. These findings support a model in which effector domains stabilize chromatin association and amplify locus-specific signal integration, thereby modulating transcriptional output beyond intrinsic DNA-binding specificity.

Building on this quantitative relationship, joint proteocistromic clustering uncovered a hierarchical organization of TFs. Cluster a TFs function as regulatory hubs characterized by extensive interaction networks, broad chromatin engagement, and enrichment in developmental and oncogenic pathways. Notably, this hierarchy was not apparent from DNA-binding motifs or domain families alone, underscoring that TF functional states arise from coordinated interaction capacity and genomic occupancy rather than intrinsic sequence recognition . This systems-level perspective explains how perturbation of a limited subset of TF hubs can disproportionately disrupt transcriptional networks in disease contexts.

Within this hierarchical landscape, TF cooperativity emerged as a central organizing principle. In vivo motif-spacing analysis identified widespread cooperative TF pairs enriched within proteocistromic cluster a, indicating that cooperative binding is a defining feature of transcriptionally dominant regulators. Functional validation of representative pairs demonstrated cooperative tumor-suppressive effects, directly linking cooperative chromatin binding to clinically relevant phenotypes. Strikingly, most cooperative pairs

were preferentially detected by BioID rather than AP-MS, suggesting that functional cooperativity frequently arises from dynamic spatial proximity within chromatin-bound assemblies rather than obligate stable dimers. This observation aligns with evidence that TF binding and function are dynamically modulated by signaling-dependent translocation and nuclear residency patterns[98]. It also parallels biochemical demonstrations that DNA sequence context and nucleosome properties, including local bendability, shape pioneer factor engagement[99]. Together, these observations reinforce that cooperative binding in vivo reflects multilayer integration of intrinsic DNA recognition, chromatin features, and interaction-dependent stabilization.

Beyond pairwise cooperativity, we identified SBSI-TF regions enriched for higher-order TF assemblies at promoters and CpG-rich loci. These regions form dense regulatory hubs distinct from canonical HOT regions[16] and are preferentially occupied by the proteocistromic cluster a TFs, suggesting that such assemblies provide robust yet adaptable transcriptional control. Structural domain-domain association analyses indicate that cooperative partnerships are constrained by complementary and shared effector architectures, providing mechanistic insight into how regulatory communities assemble at defined genomic loci.

Genetic variation provided orthogonal validation of proteocistromic states. Allele-specific binding (ASB) events were enriched among highly interactive TFs and at SBSI-TF loci, overlapping eQTLs and cancer-associated GWAS signals. Importantly, ASB burden varied substantially even among TFs within the same DNA-binding family, indicating that effector-domain interaction capacity and regulatory context, rather than motif recognition alone, govern the genetic sensitivity of chromatin engagement. These findings are consistent with our prior demonstrations that noncoding risk variants perturb lineage-restricted TF complexes, alter long-range chromatin interactions, and reshape target-gene regulation through allele-dependent binding events[100-103]. Such work established a causal framework linking regulatory variants to TF occupancy and three-dimensional genome architecture, providing a foundation for interpreting allele-specific binding within structured TF interaction hierarchies.

At the network scale, community-level analysis revealed that TFs assemble into higher-order interaction modules extensively engaged across diverse cancer systems. Communities enriched for proteocistromic TFs were preferentially associated with multiple tumor types, highlighting their central role in oncogenic signal integration. Hub TFs within these modules displayed broadly conserved expression patterns across cancers, consistent with roles in core malignant transcriptional programs rather than tumor-type-specific functions. This partitioning suggests a division between universally required transcriptional regulators and context-specific modulators, with important implications for therapeutic strategies aimed at disrupting essential oncogenic transcriptional

dependencies while preserving lineage-restricted regulatory circuits. Such stratification is consistent with emerging evidence that noncoding cis-regulatory elements and their associated TF assemblies can constitute actionable vulnerabilities within 3D genome networks in cancer[104].

Several limitations warrant consideration. Our atlas encompasses 216 TFs within a defined cellular context, whereas proteocistromic states are inherently dynamic across developmental lineages, signaling regimes, and chromatin landscapes. Although proximity labeling and AP-MS capture complementary interaction modalities, higher-order structural organization within chromatin-bound assemblies remains unresolved at molecular resolution. Emerging structural proteomics and single-molecule chromatin mapping approaches will be essential for resolving the three-dimensional architecture and temporal dynamics of transcriptional regulatory communities.

Collectively, our study establishes proteocistromics as a unified framework for decoding how TFs integrate signaling DNA-binding specificity, effector-domain interaction capacity, chromatin architecture, and genetic variation into coherent regulatory states. By revealing hierarchical TF organization, multilayer cooperativity, structurally constrained interaction communities, and genetically sensitive regulatory hotspots, we provide a mechanistic blueprint for understanding transcriptional control as an emergent systems property. Rather than viewing TFs as isolated DNA-binding entities, these results demonstrate that functional TF states arise from coordinated coupling of interaction networks and chromatin engagement. This systems-level framework offers a generalizable strategy for dissecting regulatory architecture across cell types and perturbation states, and establishes a foundation for mechanistic interrogation of transcriptional regulation at proteome-genome scale.

Materials and methods

Generation of stable Cell lines

Transcription factor (TF) open reading frame (ORF) clones were obtained as Gateway entry vectors from the Helsinki Genome Biology Unit (GBU) ORFeome Library, our previous studies,[35,39] or the DNASU plasmid repository.[108] All constructs were verified by Sanger sequencing prior to use.

Stable HEK293 cell lines were generated as previously described.[35,43] Briefly, TF ORFs were transferred into MAC-tag destination vectors using Gateway recombination cloning and introduced into Flp-In™ T-REx™ 293 host cells (Thermo Fisher Scientific) to establish isogenic, tetracycline-inducible stable cell lines. Cells were maintained in high-glucose

DMEM supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin at 37 °C in a humidified incubator with 5% CO_2.

Stable 22Rv1 prostate cancer cell lines were generated by lentiviral transduction of bait constructs containing a C-terminal FLAG-HA epitope tag and a puromycin resistance cassette under the control of the CMV promoter, as previously described.[32] Expression constructs were generated by Gateway cloning and packaged into lentiviral particles in HEK293 cells using Lipofectamine 2000 (Thermo Fisher Scientific) according to the manufacturer's instructions. Viral supernatants were applied to 22Rv1 cells, and transduced cells were selected with puromycin (2 µg/ml; Sigma-Aldrich) until stable populations were obtained. 22Rv1 cells were cultured in RPMI-1640 medium supplemented with 10% FBS and 1% penicillin-streptomycin under standard conditions (37 °C, 5% CO□).

Affinity purification and proximity-dependent biotinylation

Stable cell lines were expanded to ~80% confluence in twenty 150-mm cell culture dishes. For AP-MS, ten dishes were induced with tetracycline (2 µg/ml) for 24 hours prior to harvest. For BioID experiments, ten dishes were induced with tetracycline (2 µg/ml) and supplemented with biotin (50 µM) for 24 h before collection. Cells from five fully confluent dishes were pooled to generate one biological replicate. In total, two independent biological replicates were generated for each approach.

Cells were harvested, washed with cold PBS, and lysed on ice in lysis buffer supplemented with protease inhibitors (Sigma-Aldrich). For AP-MS, pellets were resuspended in 3 ml of ice-cold lysis buffer containing 0.5% IGEPAL CA-630, 50 mM HEPES (pH 8.0), 150 mM NaCl, 50 mM NaF, 1.5 mM NaVO3, 5 mM EDTA, 0.5 mM PMSF, and protease inhibitor cocktail. For BioID, pellets were resuspended in 3 ml of ice-cold lysis buffer containing 0.5% IGEPAL, 50 mM HEPES (pH 8.0), 150 mM NaCl, 50 mM NaF, 1.5 mM NaVO3, 5 mM EDTA, 0.1% SDS, 0.5 mM PMSF, and protease inhibitors, followed by sonication and treatment with Benzonase® nuclease (Santa Cruz Biotechnology, sc-202391) to reduce nucleic acid-mediated interactions.

Lysates were clarified by centrifugation, and supernatants were subjected to one-step affinity purification using Strep-Tactin® Sepharose® resin (IBA Lifesciences) according to established protocols. After extensive washing, bound proteins were processed for downstream mass spectrometry analysis as previously described.[43]

Sample preparation for mass spectrometry

Purified protein complexes were subjected to reduction, alkylation, and enzymatic digestion prior to mass spectrometry (MS) analysis. Proteins were reduced with 5 mM Tris(2-carboxyethyl)phosphine (TCEP; Sigma-Aldrich, C4706) at 37 °C for 20 min, followed by alkylation with 10 mM iodoacetamide (Sigma-Aldrich, Fluka 57670) for 20 min at room temperature in the dark.

Proteins were digested with 1.5 µg of sequencing-grade modified trypsin (Promega, V5111) at 37°C for 16 h. Digestion was quenched by the addition of trifluoroacetic acid (TFA) to a final concentration of 0.5% (v/v). Resulting peptides were desalted using BioPureSPN Mini C18 columns (The Nest Group) according to the manufacturer's instructions and dried under vacuum.

Dried peptides were resuspended in 30 µl of buffer A (0.1% (v/v) TFA and 1% (v/v) acetonitrile in HPLC-grade water). For LC-MS/MS analysis, 2 µl of each sample were loaded onto Evotips (Evosep) following the manufacturer's protocol.

Liquid chromatography-mass spectrometry (LC-MS/MS) analysis

Desalted peptide samples were analyzed using an Evosep One liquid chromatography system (Evosep) coupled to a trapped ion mobility quadrupole time-of-flight mass spectrometer (timsTOF Pro, Bruker Daltonics) via a CaptiveSpray nano-electrospray ion source. Peptides were separated on an 8 cm × 150 µm analytical column packed with 1.5 µm C18 particles (EV1109, Evosep) using the Evosep 60 samples per day method with a 21-min gradient.

Mobile phase A consisted of 0.1% formic acid in water, and mobile phase B consisted of 0.1% formic acid in acetonitrile. Mass spectrometry data were acquired in positive-ion mode using data-dependent acquisition (DDA) with parallel accumulation-serial fragmentation (PASEF), with a total cycle time of 0.5 s.

Raw data were processed using FragPipe v17.1 with MSFragger[109] and searched against the reviewed human UniProtKB database (downloaded on March 8, 2022). Carbamidomethylation of cysteine was specified as a fixed modification, while N-terminal acetylation, methionine oxidation, and biotinylation of lysine residues and protein N-termini were set as variable modifications. Trypsin was specified as the proteolytic enzyme, allowing up to two missed cleavages. Default instrument and label-free quantification parameters were applied. Peptide- and protein-level identifications were filtered to a false discovery rate (FDR) < 1% using Philosopher, and spectral count (SC) values from confidently identified peptides were used for downstream interaction analyses.

Statistical assessment of protein-protein interactions

High-confidence protein–protein interactions were identified using the Significance Analysis of INTeractome (SAINTexpress version 3.6.3) implemented through the ProHits-viz web platform (http://proteomics.fi/).[110] Spectral count data from biological replicates were used as input, and interactions were scored using a Bayesian framework that estimates the probability of true interaction relative to background contaminants. High-confidence interactions (HCIs) were defined based on a protein-level Bayesian false discovery rate (BFDR) ≤ 0.05.

To further reduce nonspecific background, identified interactors were additionally filtered using the Contaminant Repository for Affinity Purification-Mass Spectrometry Data (CRAPome, version 2.0). Proteins detected with a frequency ≥ 50% across CRAPome control datasets were excluded unless they exhibited an average spectral count fold change > 3 relative to controls, thereby retaining enriched, bait-specific interactors.

Chromatin immunoprecipitation sequencing (ChIP-seq)

HEK293 cells were seeded in 15-cm dishes and induced with tetracycline (final concentration, 2 µg/ml) to express the indicated TF protein for 24 h prior to harvest. ChIP-seq assays were performed as previously described, with minor modifications.[39,111] Briefly, cells were cross-linked with 1% formaldehyde for 10 min at room temperature, and the reaction was quenched with 125 mM glycine. Cells were washed twice with ice-cold PBS and lysed in hypotonic lysis buffer (20 mM Tris-Cl, pH 8.0, 10% glycerol, 10 mM KCl, 2 mM DTT, and complete protease inhibitor cocktail; Roche) to isolate nuclei. Nuclear pellets were washed once with ice-cold PBS and resuspended in a 1:1 mixture of SDS lysis buffer (50 mM Tris-HCl, pH 8.1, 1% SDS, 10 mM EDTA, and complete protease inhibitor) and ChIP dilution buffer (16.7 mM Tris-HCl, pH 8.1, 0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 167 mM NaCl, and complete protease inhibitor). Chromatin was sheared to an average fragment size of ~300 bp using a Q800R sonicator (QSonica) at 4 °C.

Dynabead protein G (Invitrogen) were pre-washed with blocking buffer (0.5% BSA in IP buffer containing 20 mM Tris-HCl, pH 8.0, 2 mM EDTA, 150 mM NaCl, 1% Triton X-100, and protease inhibitor cocktail) and incubated with anti-HA antibody (ab18181, Abcam). The chromatin lysate was then incubated with antibody-conjugated beads for 12 h at 4 °C with rotation. Beads were washed four times with wash buffer (50 mM HEPES, pH 7.6, 1 mM EDTA, 0.7% sodium deoxycholate, 1% NP-40, and 0.5 M LiCl), followed by two washes with 100 mM ammonium bicarbonate (AMBIC). DNA-protein complexes were eluted using extraction buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, and 1% SDS), and cross-links were reversed by incubation with proteinase K and NaCl. DNA was purified using the MiniElute PCR Purification Kit (Qiagen).

ChIP-seq libraries were prepared using the NEBNext® Ultra™ II DNA Library Prep Kit (E7103L, New England Biolabs) according to the manufacturer's instructions. Libraries were sequenced on an Illumina NovaSeq 6000 platform using the S4 flow cell configuration.

ChIP-seq analysis

All ChIP libraries were sequenced to generate 150 bp single-end reads. Raw sequencing quality was assessed using FastQC (v0.11.9) and summarized with MultiQC (v1.13.dev0).[112] Adapter sequences and short reads (<20 bp) were removed using Trim Galore (v0.6.7; RRID: SCR_011847). High-quality reads were aligned to the human reference genome (hg38) using Bowtie2 (v2.2.5)[113] with default parameters.

Aligned reads were filtered using SAMtools (v1.9)[114] to remove low-quality and non-informative alignments, including: (i) reads with mapping quality <30, (ii) unmapped reads, (iii) non-primary alignments, (iv) reads failing platform/vendor quality checks, (v) PCR or optical duplicates, and (vi) supplementary alignments, using the parameters "-q 30 -F 3844". In addition, duplicate reads were further marked and removed using Picard MarkDuplicates (v2.25.1; RRID: SCR_006525).

Peak calling was performed using MACS2 (v2.1.4)[115] with a significance threshold of $p < 1$ x 10□□. Identified peaks were annotated to genomic features and nearby genes using the HOMER script **annotatePeaks.pl** (v4.11.1).[116]

Motif analysis

De novo motif discovery was performed using the MEME-ChIP suite (v5.0.5).[117] Peak summits were extended to 200 bp windows (±100 bp from the summit), as previously described.[45,46] Peaks overlapping ENCODE blacklisted regions (ENCODE v4 GRCh38 blacklist; ENCODE accession: ENCFF356LFX) were excluded using the **subtract** function in BEDTools (v2.30.0), allowing a maximum overlap fraction of 25% between peaks and blacklist regions (-f 0.25).

Filtered peaks were ranked by peak-calling *p* values, and the top 500 peaks were selected. Corresponding genomic sequences were extracted from the hg38 reference genome to generate FASTA files for motif analysis. Identified motifs were compared against curated motif databases, including HOCOMOCO v11, JASPAR 2022 CORE, and the Jolma et al. 2013 collection[47], and the top five significantly enriched motifs were reported.

Cobinding and tethered binding identification

As previously described,[46] two non-mutually exclusive binding modes—cobinding and tethered binding, were evaluated to explain the occurrence of secondary motifs within ChIP-seq peaks. In this framework, the primary motif was defined as the most significantly enriched motif identified by MEME-ChIP, whereas secondary motifs refer to the remaining significantly enriched motifs. Cobinding denotes direct and independent binding of two transcription factors (TF1 and TF2) to adjacent DNA sites, whereas tethered binding indicates that TF1 is recruited to chromatin through protein–protein interaction with TF2, which directly binds DNA.[46]

To distinguish between these two binding modes, we applied the quantitative criteria proposed by Yu et al.[45] Specifically, three fractions were calculated:

X, the fraction of peaks containing only the canonical (primary) motif;

Y, the fraction of peaks containing only the non-canonical (secondary) motif;

Z, the fraction of peaks containing both motifs.

Binding mode was inferred as follows: (i) if $\mathbf{Y \geq X}$ and $\mathbf{Y \geq Z}$, tethered binding was inferred; (ii) if $\mathbf{Y < X}$ or $\mathbf{Y < Z}$, cobinding was inferred.

Because transcription factors within the same family often share highly similar DNA-binding domains (DBDs) and recognize related core motifs, motif relationships were not inferred between TFs belonging to the same DBD family. Instead, TFs sharing the same DBD were grouped into a single family,[2] and the canonical motif for a given TF was defined as the motif bound by that TF or by TFs sharing the same DBD.

Motif occurrences were identified using FIMO (v5.0.5) [118] by scanning the top 500 ChIP-seq peaks, with a significance threshold of $p < 1e\text{-}4$, consistent with prior studies.[45] The resulting motif counts were used to calculate X, Y, and Z values for binding mode classification.

TF categorization based on proteomics

To improve the robustness and generalizability of TF categorization, we applied the following criteria and analytical strategies. First, only TFs that were profiled as baits in both AP-MS and BioID experiments were included in the analysis (n = 195). Notably, inclusion was based on the availability of bait measurements in both platforms, rather than requiring that individual PPIs be detected by both methods. Second, the ED activity of each bait TF was approximated by the total number of interacting prey TFs and cofactor proteins identified across experiments. Third, rather than stratifying prey proteins by

functional class, all prey proteins were analyzed collectively to derive a global interaction burden per TF.

Unsupervised clustering was performed using the k-means algorithm, with the number of clusters set to three, to classify TFs into groups with high (P1), intermediate (P2), and low (P3) levels of associated prey proteins. Although silhouette coefficient analysis suggested two clusters as the statistically optimal solution for both k-means and self-organizing map (SOM) approaches, we selected three clusters for k-means to better reflect the graded nature of TF regulatory activity. The intermediate cluster (P2) captured TFs with moderate cofactor interaction profiles, which are supported by prior studies and by our integrated proteocistromic observations. Comparative evaluation across clustering methods and cluster numbers indicated that k-means with n = 3 yielded the most distinct and biologically interpretable groupings, whereas n = 2 failed to resolve intermediate states, and higher cluster numbers, particularly SOM with n = 4, produced overlapping or disordered partitions consistent with overfitting. Therefore, k-means clustering with three groups was selected as a balanced and biologically motivated classification strategy.

To investigate cistromic features associated with the proteomics-derived TF groups, we quantified the number of ChIP-seq peaks for TFs in each group (P1–P3); TFs lacking cistromic data were excluded from this analysis. Two-tailed t tests were used to compare mean peak numbers between groups. Peaks were further stratified by fold enrichment over input, as determined by MACS2, into all peaks (fold enrichment ≥ 1), high-confidence peaks (≥ 5-fold), and very high-confidence peaks (≥ 10-fold). In addition, peaks from TFs within each group were merged, and signal profiles around transcription start sites (TSSs) were computed using computeMatrix and visualized using plotHeatmap in deepTools (v3.5.2),[119] across ±3 kb of TSSs with a bin size of 50 bp. TSS coordinates for the hg38 genome assembly were obtained from the UCSC Genome Browser.

TF categorization based on cistromics

Enrichment of transcription factor (TF) binding sites within defined epigenomic regulatory elements provides insights into their regulatory modes and functional potential. Genome-wide regulatory regions were obtained from the Ensembl Regulatory Build,[120] using the summary track that integrates regulatory states across multiple cell types, including promoters, enhancers, transcription factor binding sites, CTCF binding sites, and open chromatin regions. A TF peak was considered enriched in a regulatory category if at least one base pair overlapped with the corresponding regulatory annotation.

For each TF, the number of peaks overlapping each regulatory category was calculated, and regulatory occupancy was defined as the ratio of peaks overlapping a given regulatory element to the total number of peaks for that TF. TFs with available cistromic data (n =

206) were clustered based on their occupancy profiles across regulatory categories using hierarchical clustering with complete linkage, implemented via the hclust function and visualized with the ComplexHeatmap package (v2.6.2) in R.

Based on occupancy patterns at promoter, enhancer, and CTCF-binding regions, TFs were classified into five distinct groups: C1, predominantly promoter-bound; C2, predominantly enhancer-bound; C5, predominantly CTCF-associated; C3, displaying balanced occupancy across promoters, enhancers, and CTCF sites; and C4, exhibiting mixed or atypical binding profiles not conforming to the other categories. These classifications reflect differential tendencies toward proximal versus distal gene regulation.

To investigate proteomic correlates of cistromic TF classes, we quantified the number of protein-protein interactions (PPIs) detected for TFs in each category (assigning a value of zero to TFs without detected interactions). Pairwise comparisons of mean PPI counts between TF categories were performed using two-tailed *t* tests.

Proteocistromic TF clustering

Proteocistromic features were designed to integratively quantify the regulatory impact of TFs by combining proteomic and cistromic measurements. Specifically, the number of TF-TF and TF-cofactor PPIs as a metric was used to assess the interaction-mediated regulatory potential of each TF. In parallel, cistromic influence was evaluated using the number of high-confidence genomic binding sites (defined as ≥10-fold enriched peaks) and the number of peaks colocalizing with key regulatory elements, including promoters, enhancers, and CTCF-binding sites, to represent DBD-mediated regulatory capacity.

For clustering analysis, proteocistromic feature values from TFs with complete datasets (n = 184) were log10-transformed to reduce skewness and improve comparability across features. The optimal number of clusters was determined by calculating silhouette scores for k values ranging from 2 to 10 using k-means clustering, and the k value that maximized the silhouette score was selected. Based on this criterion, k = 3 was chosen and used to partition TFs into three distinct clusters.

Clustering results were visualized using parallel coordinates plots generated with the ggparcoord function from the GGally package (v2.2.0), enabling simultaneous inspection of feature distributions across clusters. In addition, z-score normalization was applied to individual proteocistromic features, and heatmaps were generated to assess relative feature abundance within and between clusters. A composite proteocistromic merged score, defined as the sum of normalized feature values across all proteocistromic dimensions, was further calculated to represent the overall regulatory potential of each TF.

Based on proteocistromic profiles, TFs in Cluster a were classified as relatively active, whereas TFs in Clusters b and c were considered relatively inactive. TFs from the active and inactive groups were subjected to pathway enrichment analysis using Metascape[121]. To visualize functional relationships among enriched pathway, a subset of terms was rendered as a network in Cytoscape (v3.10.1), in which nodes represent enriched terms and edges connect terms with similarity scores >0.3, as defined by Metascape. Nodes were colored by cluster identity to indicate functional proximity among related biological processes.

Pfam domain mapping and domain-domain interaction analysis

All bait and prey proteins was mapped to were annotated with Pfam[122] domains using the current Pfam database release (last modified: 2023-01-03; https://pfam-docs.readthedocs.io/en/latest/ftp-site.html). For proteins containing multiple instances of the same Pfam domain, duplicate domains were collapsed such that each unique domain was counted only once per protein, thereby preventing inflation of domain–domain interaction frequencies and ensuring that interactions were defined between distinct domain types.[33]

To identify statistically significant domain–domain associations, Fisher's exact test was applied to evaluate whether pairs of Pfam domains co-occurred in interacting bait–prey protein pairs more frequently than expected by chance[33]. Resulting *p* values were adjusted for multiple testing using the Benjamini-Hochberg procedure to control the false discovery rate (FDR).

TF pair (motif spacing) analysis

To characterize spatial relationships between TF binding motifs, we implemented the TF spacing analysis pipeline described by Shen et al.[71] using the publicly available workflow (https://github.com/zeyang-shen/spacing_pipeline). For each TF ChIP-seq dataset, peaks were re-called using HOMER (v4.11.1)[116] with the 'findPeaks' function and parameters '-style factor -fdr 0.001 -size 200' to obtain high-confidence binding regions.

Position weight matrices (PWMs) for TFs were obtained from the JASPAR 2022 database[123] and the SCENIC+ clustered motif collection [124] (https://resources.aertslab.org/cistarget/motif_collections/). For TFs lacking annotated motifs in both databases, de novo motif discovery was performed using HOMER. Following motif compilation, the spacing pipeline was applied to quantify and classify spacing relationships between TF motif pairs across binding regions.

Significant spacing relationships were determined using the Kolmogorov-Smirnov (KS) test, which evaluates deviations from uniform spacing distributions. A stringent

significance threshold of $p < 6.25 \times 10^{-6}$ was applied, corresponding to a family-wise error rate of 0.05 corrected for 200 TF pairs and four spacing modes (0.05/200/4). TF pairs were classified as exhibiting relaxed spacing when constrained spacing values were reported as “NaN” by the pipeline, whereas pairs with defined constrained distances were classified as constrained spacing interactions.

To investigate the clinical relevance of TF pairs, gene expression (TPM) and overall survival data across 33 cancer types were obtained from The Cancer Genome Atlas (TCGA) via the cBioPortal platform[125] (https://www.cbioportal.org/). Kaplan-Meier survival analyses were performed using the survival package (v3.2.10) in R for TF pairs with the highest proteocistromic scores in both relaxed and constrained spacing categories, specifically MAX-KLF5 and FOXJ2-KLF4. Patients were stratified into four groups based on TF expression levels: both low, TF1 low/TF2 high, TF1 high/TF2 low, and both high. Low and high expression were defined relative to the cohort-specific mean expression of each gene. Survival differences were evaluated using the log-rank test, with $p < 0.05$ considered statistically significant.

Previous studies have identified extensive TF-TF motif interactions using CAP-SELEX assays.[69] To validate TF pairs identified in this study, we compared them with CAP-SELEX-derived TF pairs, as well as with TF pairs supported by protein–protein interaction data, independent of motif orientation. Motif enrichment validation was performed for TF pairs detected in both datasets. For each TF, the top 500 significant ChIP-seq peaks were selected and analyzed using CentriMo[126] to assess enrichment of individual TF motifs and paired motifs, with statistical significance defined by adjusted *E* values < 0.05. PWMs for individual TF motifs were obtained from JASPAR, while PWMs for TF-pair motifs were derived from the CAP-SELEX dataset.[69]

Lentiviral constructs, virus production, and transduction

Transcription factor (TF) coding sequences were cloned into a lentiviral destination vector containing a C-terminal FLAG-HA epitope tag and a puromycin resistance cassette driven by the CMV promoter,[32] using Gateway recombination cloning. To generate stable TF-overexpressing 786-O, ACHN, and Caki-1 cell lines, third-generation lentiviral particles were produced in HEK293 cells.

Briefly, HEK293 cells were seeded in 6-cm dishes and transfected at 60–80% confluence using polyethyleneimine (PEI). Cells were co-transfected with 3 µg of TF expression plasmid together with 1 µg each of the packaging plasmids pMDLg/pRRE, pRSV-Rev, and the envelope plasmid pVSV-G. Culture medium was replaced with fresh complete medium 6 h post-transfection. Viral supernatants were collected 48 h after transfection and used directly for infection or filtered through a 0.45-µm filter to remove cell debris.

For transduction, 786-O, ACHN, and Caki-1 cells were plated in 6-well plates and infected at 60–70% confluence with lentiviral supernatant for 24 h. Following infection, cells were selected with puromycin (1 μg/mL; Sigma) to establish stable overexpression cell lines.

Migration assays

Cell migration was assessed using Transwell chambers (8-μm pore size; Cat. No. 353097, BD Biosciences). 786-O, ACHN, and Caki-1 cells were harvested and resuspended in serum-free medium, and 200 μL of cell suspension was added to the upper chamber of each insert. The lower chambers were filled with complete medium as a chemoattractant. After 24 h of incubation at 37 °C, non-migrated cells on the upper surface of the membrane were removed using a cotton swab. Migrated cells on the lower surface were fixed with 4% paraformaldehyde and stained with crystal violet staining solution (Beyotime). Cells were imaged and quantified by counting in multiple random fields per insert under a light microscope. Data were obtained from three independent inserts per condition, and statistical significance was evaluated using two-tailed Student's *t* tests.

Colony formation assay

For clonogenic assays, 786-O, ACHN, and Caki-1 cells were seeded at a density of 250 cells per well in 12-well plates and cultured for 10–14 days at 37 °C in a humidified incubator with 5% CO□. Culture medium was replaced every 3–4 days. At the end of the incubation period, colonies were fixed with 1 mL of 4% paraformaldehyde per well and stained with crystal violet for 20 min. Wells were rinsed gently with water to remove excess stain and air-dried before imaging. Colonies were photographed and, where indicated, quantified using standard image analysis procedures.

Allele-specific binding (ASB) analysis

Allele-specific binding (ASB) events were identified using BaalChIP,[83] which applies a Bayesian statistical framework to detect transcription factor (TF) binding biases caused by single-nucleotide variants across the genome. Stringent filtering and quality control (QC) procedures were performed according to the BaalChIP pipeline. Briefly, heterozygous single-nucleotide polymorphisms (SNPs) were obtained from the dbSNP database and provided as the hets file, while corresponding ChIP-seq input data for each TF were used as genomic DNA (gDNA) controls. Allele-specific read counts were calculated using default BaalChIP parameters, and only heterozygous SNPs overlapping TF ChIP-seq peak regions were retained for further analysis.

Several sources of technical bias were sequentially corrected during QC. First, SNPs located within ENCODE blacklisted regions or collapsed repeat regions were excluded.[116] Second, SNPs exhibiting intrinsic allelic bias were filtered out.[128,129] In addition, potential

homozygous SNPs were removed, and both reference mapping bias and reference allele frequency bias were explicitly modeled and corrected within the BaalChIP framework. Final ASB calls were extracted, and the logical ASB status reported by BaalChIP was used for downstream analyses.

To investigate the potential functional relevance of ASB events, we performed colocalization analyses with cis-expression quantitative trait loci (cis-eQTLs) in both tumor and normal tissues and summarized overlaps between ASB SNPs and reported GWAS variants. Tumor cis-eQTLs were obtained from the PancanQTL database,[130] (http://gong_lab.hzau.edu.cn/PancanQTL/), which provides data across 33 cancer types. Normal tissue cis-eQTLs were derived from the GTEx v8 release,[131] (https://gtexportal.org/home/), encompassing 49 tissue types. GWAS-associated SNPs were retrieved from the GWAS Catalog,[132] including all entries available up to December 20, 2023 (https://www.ebi.ac.uk/gwas/). Genome-wide annotation of ASB loci relative to genomic features and nearby genes was performed using the annotatePeaks.pl script from the HOMER software.

IP-SNP-seq validation of allele-specific binding events

To experimentally validate allele-specific binding (ASB) events, we selected the top 200 ASB SNPs from each cluster, ranked by total read counts, yielding 1,145 unique SNPs for validation. For these loci, a total of 4,660 single-stranded oligonucleotides were designed, including 20 positive and 20 negative control sequences. For each SNP, four 21-nt single-stranded oligonucleotides with the SNP centrally positioned were synthesized (Tsingke Biotech, China), each provided at 20 μM in 25 μL.

To generate double-stranded (ds) oligonucleotides, 6 μL of each complementary forward and reverse strand were mixed with 3 μL of 5× annealing buffer, denatured at 95 °C for 3 min, and annealed by gradual cooling from 95 °C to 25 °C over 70 min. Resulting ds-oligos were pooled in equimolar amounts to generate a master oligo pool representing all selected SNP sequences.

Nuclear extracts were prepared from HEK293 and prostate cancer 22Rv1 cells using NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Fisher Scientific), and protein concentrations were determined using a BCA protein assay kit (P0011, Beyotime). Aliquots (25 μL) were stored at −80 °C until use. For DNA-protein binding reactions, 5 μL of the ds-oligo pool was incubated with 10 μg nuclear extract in 1× TF binding buffer (Signosis) at 23 °C for 30 min. Complexes were captured using isolation columns, washed five times, and protein-bound ds-oligos were eluted with Elution Buffer, followed by purification using the Oligo Clean & Concentrator kit (Zymo Research). All assays were performed in

duplicate. Purified oligo concentrations were measured using the NGS dsDNA HS assay kit (FP002, ABP Biosciences).

For library preparation, 2 ng of purified ds-oligos were used as input for the ThruPLEX DNA-seq kit (Takara Biomedical Technology) with 12 PCR cycles. Libraries were purified using Agencourt AMPure XP beads (Beckman Coulter), quantified with a Qubit fluorometer, pooled, and sequenced on an Illumina NovaSeq platform (GENEWIZ, China) with 150-bp paired-end reads.

Sequencing reads were first subjected to quality control using FastQC (v0.11.9), and adapters and low-quality reads were removed using Trim Galore (v0.6.7). Trimmed reads were aligned to a custom reference containing 4,580 unique oligonucleotide sequences representing 1,145 SNP loci using Bowtie (v1.3.1) with parameters “-m 1 -v 0” to retain only uniquely and perfectly matched reads. Read counts for reference and variant alleles were quantified separately for both immunoprecipitated (test) and input samples.

Allele binding differences were quantified using a Binding Allele Bias (BAB) score defined by the formula: BAB score = log□ [(test_RCvariant / test_RCreference) / (input_RCvariant / input_RCreference)], where RC denotes read counts for each allele. An absolute BAB score ≥ 0.58, corresponding to a ≥ 1.5-fold allelic imbalance, was considered indicative of significant allele-specific binding.

Protein community analysis

Protein interaction communities were identified using the Markov Cluster (MCL) algorithm,[90] implemented with MCL software (v22-282; https://epubs.siam.org/doi/10.1137/040608635). All bait-prey interactions were formatted as edge lists, with each interaction represented by a peptide spectral match (PSM) value as edge weight. Clustering was performed using an inflation parameter of 2, and clusters containing fewer than three proteins were excluded from downstream analyses.

To assess inter-community connectivity, the number of protein-protein interactions within and between clusters was quantified. Enrichment of interactions between cluster pairs was evaluated using Fisher’s exact test with multiple-testing correction, and cluster pairs with false discovery rate (FDR) < 0.05 were considered significantly interconnected.[34]

To further characterize the nature of connectivity between two communities (community 1, C1; community 2, C2), interactions were classified into three categories: **Both**, PPIs involving proteins from both C1 and C2; **Just_C1**, PPIs involving only proteins within C1; and **Just_C2**, PPIs involving only proteins within C2. A pair of communities was defined as exhibiting **close interaction** if the number of *Both* interactions exceeded the number of either *Just_C1* or *Just_C2* interactions; otherwise, the communities were considered

relatively isolated. This framework enabled discrimination between functionally integrated versus modular community structures within the protein interaction network.

Community-cancer interaction analysis

Cancer-associated genes were obtained from CancerMine,[91] a literature-curated resource cataloging oncogenes, tumor suppressors, and cancer driver genes (downloaded on 2023-11-30). Cancer types were grouped according to human organ systems. To evaluate enrichment of cancer associations within protein communities, we considered four parameters: (i) the number of cancer types represented in each community, (ii) the number of cancer-associated genes within each community, (iii) the number of communities associated with each cancer type, and (iv) the total size of the community–cancer interaction network. Enrichment significance was assessed using hypergeometric tests, followed by multiple-testing correction using the Benjamini–Hochberg procedure, with FDR < 0.05 considered statistically significant.

To further assess the pan-cancer relevance of proteocistromic transcription factors (TFs) influenced by community interactions, we performed dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) on TF expression profiles across multiple cancer types. TFs from closely interacting communities were compared with other proteocistromic TFs. UMAP was performed with parameters *n_neighbors* = 30 and *min_dist* = 0.3. Gene expression data, expressed as log□(TPM + 0.001), were obtained from the UCSC Toil RNA-seq Recompute Compendium[133] via UCSC Xena (https://xena.ucsc.edu/). This dataset is processed using a uniform pipeline and is largely free of batch effects, enabling reliable cross-cancer clustering analyses.

Declarations

Acknowledgements

This work was supported by the Shanghai Interactional Collaborative Project (23410713300); the Shenzhen Medical Research Fund (SMRF) (C2503001); the National Natural Science Foundation of China (82073082; 82311530050; 82203416); the Jane and Aatos Erkko Foundation; the Finnish Cancer Foundation; the Sigrid Juseliuksen Saatio; the Thelma Mäkikyrö Foundation; the University of Oulu Profi6 Fibrobesity programme; and the Research Council of Finland (Decision No. 336449). High-performance computing resources were provided by CSC-IT Center for Science Ltd, and were also supported by the Medical Research Data Center of Shanghai Medical College of Fudan University, as well as by the High-Performance Computing Platform of Suzhou Institute of Systems Medicine□Chinese Academy of Medical Sciences & Peking Union Medical College.

Conflict of interests

The authors declare no competing interests.

Contributions

Conceptualization, G.-H.W., M.V.; Z.T. performed the experiments. Z.W. designed, performed, and interpreted most of bioinformatics analysis. X.L. designed and performed mass spectrometry analyses. Z.W., Z.T., X.L., G.D., and G.-H.W. prepared figures. Methodology, Z.W., Z.T., X.L., G.D., B.L., Z.P., W.X., L.Q., Y.Y., S.M., X.Y., Q.Z., A.M., M.V., and G.-H.W.; Software, Z.W., Z.T., X.L., G.-H.W.; Validation, Z.W., X.L., G.D., Z.P., W.X.; Formal analysis, Z.W., X.L., M.V., and G.-H.W.; Investigation, Z.T., X.L., B.L., G.D., M.V., and G.-H.W.; Resources, M.V., G.-H.W.; Data curation, Z.W., Z.T., X.L., M.V., G.-H.W.; Writing – original draft, Z.W., Z.T., X.L., G.-H.W.; Writing – review & editing, Z.W., Z.T., X.L., G.D., M.V., G.-H.W. with inputs from all authors; Visualization, Z.W., X.L., M.V., G.-H.W.; Supervision, M.V., G.-H.W.; Project administration, A.M., M.V., G.-H.W.; Funding acquisition, M.V., G.-H.W.

Supplemental Informations

Figures S1-S9 and Tables S1-S7.

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Gong-Hong Wei (gonghong_wei@fudan.edu.cn).

Data and code availability

All raw and processed sequencing data generated in this study have been deposited in the Gene Expression Omnibus (GEO) and are publicly available as of the date of publication under accession number GSE271763 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE271763). During peer review, access to the private record was provided via the secure token: cvgbaywubtmxdqf.

Protein–protein interaction (PPI) source data underlying the figures are provided with this paper in **Table S2**.

All original code used in this study has been deposited in GitHub and is available at https://github.com/zxwang-cloak/Proteocistromic

Mass spectrometry (MS) proteomics data have been deposited in the MassIVE repository with web access under accession number MSV000093138 (doi:10.25345/C5RX93Q4N]).

Any additional information required to reanalyze the data reported in this paper is available from the Lead Contact upon reasonable request. Requests for resources and reagents should also be directed to and will be fulfilled by the Lead Contact.

References

1. Brivanlou, A.H. & Darnell, J.E., Jr. Signal transduction and the control of gene expression. *Science* **295**, 813-8 (2002).
2. Lambert, S.A. et al. The Human Transcription Factors. *Cell* **175**, 598-599 (2018).
3. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252-63 (2009).
4. Wei, G.H., Liu, D.P. & Liang, C.C. Charting gene regulatory networks: strategies, challenges and perspectives. *Biochem J* **381**, 1-12 (2004).
5. Wang, M. et al. Integrative analysis of non-small cell lung cancer identifies Jumonji domain-containing 6/ETS homologous factor axis as a target to overcome radioresistance. *Signal Transduct Target Ther* **10**, 391 (2025).
6. Jang, M., Choi, H.J., Lee, H.J. & Kim, H.N. Hypoglycemia induces brain metabolic reprogramming and neurodegeneration via serum response factor and myocardin-related transcription factor-A. *Signal Transduct Target Ther* **10**, 412 (2025).
7. Darnell, J.E., Jr. Transcription factors as targets for cancer therapy. *Nat Rev Cancer* **2**, 740-9 (2002).
8. Bushweller, J.H. Targeting transcription factors in cancer - from undruggable to reality. *Nat Rev Cancer* **19**, 611-624 (2019).
9. Chen, A. & Koehler, A.N. Transcription Factor Inhibition: Lessons Learned and Emerging Targets. *Trends Mol Med* **26**, 508-518 (2020).
10. Basu, S. et al. Rational optimization of a transcription factor activation domain inhibitor. *Nat Struct Mol Biol* **30**, 1958-1969 (2023).
11. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-86 (2014).
12. Meyer, K.D., Lin, S.C., Bernecky, C., Gao, Y. & Taatjes, D.J. p53 activates transcription by directing structural shifts in Mediator. *Nat Struct Mol Biol* **17**, 753-60 (2010).
13. Vojnic, E. et al. Structure and VP16 binding of the Mediator Med25 activator interaction domain. *Nat Struct Mol Biol* **18**, 404-9 (2011).
14. Soto, L.F. et al. Compendium of human transcription factor effector domains. *Mol Cell* **82**, 514-526 (2022).
15. Neph, S. et al. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274-86 (2012).

16. Partridge, E.C. et al. Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature* **583**, 720-728 (2020).
17. Ravasi, T. et al. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744-52 (2010).
18. Lambert, S.A. et al. The Human Transcription Factors. *Cell* **172**, 650-665 (2018).
19. Tsai, K.L. et al. A conserved Mediator-CDK8 kinase module association regulates Mediator-RNA polymerase II interaction. *Nat Struct Mol Biol* **20**, 611-9 (2013).
20. Kim, M. et al. A protein interaction landscape of breast cancer. *Science* **374**, eabf3066 (2021).
21. Zheng, F. et al. Interpretation of cancer mutations using a multiscale map of protein systems. *Science* **374**, eabf3067 (2021).
22. Näär, A.M., Lemon, B.D. & Tjian, R. Transcriptional coactivator complexes. *Annu Rev Biochem* **70**, 475-501 (2001).
23. Josefowicz, S.Z. et al. Chromatin Kinases Act on Transcription Factors and Histone Tails in Regulation of Inducible Transcription. *Mol Cell* **64**, 347-361 (2016).
24. Mark, K.G. & Rape, M. Ubiquitin-dependent regulation of transcription in development and disease. *EMBO Rep* **22**, e51078 (2021).
25. Mouchiroud, L., Eichner, L.J., Shaw, R.J. & Auwerx, J. Transcriptional coregulators: fine-tuning metabolism. *Cell Metab* **20**, 26-40 (2014).
26. Donner, A.J., Ebmeier, C.C., Taatjes, D.J. & Espinosa, J.M. CDK8 is a positive regulator of transcriptional elongation within the serum response network. *Nat Struct Mol Biol* **17**, 194-201 (2010).
27. Boija, A. et al. Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842-1855.e16 (2018).
28. Sabari, B.R. et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**(2018).
29. Nair, S.J. et al. Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. *Nat Struct Mol Biol* **26**, 193-203 (2019).
30. Boehning, M. et al. RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nat Struct Mol Biol* **25**, 833-840 (2018).
31. Tycko, J. et al. High-Throughput Discovery and Characterization of Human Transcriptional Effectors. *Cell* **183**, 2020-2035 e16 (2020).
32. Huttlin, E.L. et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* **184**, 3022-3040 e28 (2021).
33. Huttlin, E.L. et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425-440 (2015).

34. Huttlin, E.L. et al. Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505-509 (2017).
35. Goos, H. et al. Human transcription factor protein interaction networks. *Nat Commun* **13**, 766 (2022).
36. Göös, H. et al. Human transcription factor protein interaction networks. *Nat Commun* **13**, 766 (2022).
37. Wang, C.I. et al. Chromatin proteins captured by ChIP-mass spectrometry are linked to dosage compensation in Drosophila. *Nat Struct Mol Biol* **20**, 202-9 (2013).
38. Liu, X. et al. An AP-MS- and BioID-compatible MAC-tag enables comprehensive mapping of protein interactions and subcellular localizations. *Nat Commun* **9**, 1188 (2018).
39. Gawriyski, L. et al. Interaction network of human early embryonic transcription factors. *EMBO Rep* **25**, 1589-1622 (2024).
40. Pinero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* **48**, D845-D855 (2020).
41. Xie, G. et al. MLL3/MLL4 methyltransferase activities control early embryonic development and embryonic stem cell differentiation in a lineage-selective manner. *Nat Genet* **55**, 693-705 (2023).
42. Froimchuk, E., Jang, Y. & Ge, K. Histone H3 lysine 4 methyltransferase KMT2D. *Gene* **627**, 337-342 (2017).
43. Liu, X., Salokas, K., Weldatsadik, R.G., Gawriyski, L. & Varjosalo, M. Combined proximity labeling and affinity purification-mass spectrometry workflow for mapping and visualizing protein interaction networks. *Nat Protoc* **15**, 3182-3211 (2020).
44. Go, C.D. et al. A proximity-dependent biotinylation map of a human cell. *Nature* **595**, 120-124 (2021).
45. Yu, C.P. et al. Discovering unknown human and mouse transcription factor binding sites and their characteristics from ChIP-seq data. *Proc Natl Acad Sci U S A* **118**(2021).
46. Wang, J. et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* **22**, 1798-812 (2012).
47. Jolma, A. et al. DNA-binding specificities of human transcription factors. *Cell* **152**, 327-39 (2013).
48. Wei, G.H. et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J* **29**, 2147-60 (2010).
49. Weirauch, M.T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).

50. Gunther, M., Laithier, M. & Brison, O. A set of proteins interacting with transcription factor Sp1 identified in a two-hybrid screening. *Mol Cell Biochem* **210**, 131-42 (2000).
51. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402-408 (2020).
52. Grandori, C., Cowley, S.M., James, L.P. & Eisenman, R.N. The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu Rev Cell Dev Biol* **16**, 653-99 (2000).
53. Del Toro, N. et al. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res* **50**, D648-D653 (2022).
54. Alerasool, N., Leng, H., Lin, Z.Y., Gingras, A.C. & Taipale, M. Identification and functional characterization of transcriptional activators in human cells. *Mol Cell* **82**, 677-695 e7 (2022).
55. Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Curr Opin Genet Dev* **43**, 73-81 (2017).
56. Wang, Z. et al. Interplay between cofactors and transcription factors in hematopoiesis and hematological malignancies. *Signal Transduct Target Ther* **6**, 24 (2021).
57. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol Cell* **83**, 373-392 (2023).
58. Kim, S. et al. DNA-guided transcription factor cooperativity shapes face and limb mesenchyme. *Cell* **187**, 692-711 e26 (2024).
59. Bell, C.C. et al. Comparative cofactor screens show the influence of transactivation domains and core promoters on the mechanisms of transcription. *Nat Genet* (2024).
60. DelRosso, N. et al. Large-scale mapping and mutagenesis of human transcriptional effector domains. *Nature* **616**, 365-372 (2023).
61. UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523-D531 (2023).
62. Whitfield, T.W. et al. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol* **13**, R50 (2012).
63. MacQuarrie, K.L., Fong, A.P., Morse, R.H. & Tapscott, S.J. Genome-wide transcription factor binding: beyond direct target regulation. *Trends Genet* **27**, 141-8 (2011).
64. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
65. Gronostajski, R.M. Roles of the NFI/CTF gene family in transcription and development. *Gene* **249**, 31-45 (2000).
66. Luna-Zurita, L. et al. Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. *Cell* **164**, 999-1014 (2016).

67. Morgunova, E. & Taipale, J. Structural perspective of cooperative transcription factor binding. *Curr Opin Struct Biol* **47**, 1-8 (2017).
68. Wang, Z., Wu, Y., Li, L. & Su, X.D. Intermolecular recognition revealed by the complex structure of human CLOCK-BMAL1 basic helix-loop-helix domains with E-box DNA. *Cell Res* **23**, 213-24 (2013).
69. Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384-8 (2015).
70. Mirny, L.A. Nucleosome-mediated cooperativity between transcription factors. *Proc Natl Acad Sci U S A* **107**, 22534-9 (2010).
71. Shen, Z. et al. Systematic analysis of naturally occurring insertions and deletions that alter transcription factor spacing identifies tolerant and sensitive transcription factor pairs. *Elife* **11**(2022).
72. Kim, D.I. et al. Probing nuclear pore complex architecture with proximity-dependent biotinylation. *Proc Natl Acad Sci U S A* **111**, E2453-61 (2014).
73. Roux, K.J., Kim, D.I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J Cell Biol* **196**, 801-10 (2012).
74. Karlseder, J., Rotheneder, H. & Wintersberger, E. Interaction of Sp1 with the growth- and cell cycle-regulated transcription factor E2F. *Mol Cell Biol* **16**, 1659-67 (1996).
75. Marais, R., Wynne, J. & Treisman, R. The SRF accessory protein Elk-1 contains a growth factor-regulated transcriptional activation domain. *Cell* **73**, 381-93 (1993).
76. Nagai, N. et al. Downregulation of ERG and FLI1 expression in endothelial cells triggers endothelial-to-mesenchymal transition. *PLoS Genet* **14**, e1007826 (2018).
77. O'Shea, E.K., Rutkowski, R. & Kim, P.S. Mechanism of specificity in the Fos-Jun oncoprotein heterodimer. *Cell* **68**, 699-708 (1992).
78. Kong, S.L., Li, G., Loh, S.L., Sung, W.K. & Liu, E.T. Cellular reprogramming by the conjoint action of ERalpha, FOXA1, and GATA3 to a ligand-inducible growth state. *Mol Syst Biol* **7**, 526 (2011).
79. Yan, J. et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801-13 (2013).
80. Wagner, A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **15**, 776-84 (1999).
81. Wreczycka, K. et al. HOT or not: examining the basis of high-occupancy target regions. *Nucleic Acids Res* **47**, 5735-5745 (2019).
82. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427-D432 (2019).

83. de Santiago, I. et al. BaalChIP: Bayesian analysis of allele-specific transcription factor binding in cancer genomes. *Genome Biol* **18**, 39 (2017).
84. Hallikas, O. et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47-59 (2006).
85. Bhagwat, A.S. & Vakoc, C.R. Targeting Transcription Factors in Cancer. *Trends Cancer* **1**, 53-65 (2015).
86. Menche, J. et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).
87. Boulay, G. et al. Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain. *Cell* **171**, 163-178 e19 (2017).
88. Creixell, P., Schoof, E.M., Erler, J.T. & Linding, R. Navigating cancer network attractors for tumor-specific therapy. *Nat Biotechnol* **30**, 842-8 (2012).
89. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
90. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-84 (2002).
91. Lever, J., Zhao, E.Y., Grewal, J., Jones, M.R. & Jones, S.J.M. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat Methods* **16**, 505-507 (2019).
92. Lee, T.I. & Young, R.A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237-51 (2013).
93. Henley, M.J. & Koehler, A.N. Advances in targeting 'undruggable' transcription factors with small molecules. *Nat Rev Drug Discov* **20**, 669-688 (2021).
94. Badis, G et al. Diversity and complexity in DNA recognition by transcription factors. Science 324, 1720-1723 (2009).
95. Barrera, L.A. et al. Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science 351, 1450-1454 (2016).
96. Brand, M et al. Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics. Nat Struct Mol Biol 11, 73-80 (2004).
97. Gibson, T.J. et al. Protein-intrinsic properties and context-dependent effects regulate pioneer factor binding and function. Nat Struct Mol Biol 31, 548-558 (2024).
98. Hao N, and O'Shea EK. Signal-dependent dynamics of transcription factor translocation controls gene expression. Nat Struct Mol Biol 19, 31-39 (2011).
99. Mariani L, Liu X, Lee K, Gisselbrecht SS, Cole PA, Bulyk ML. DNA bendability regulates transcription factor binding to nucleosomes. Nat Struct Mol Biol 32, 2185-2195 (2025).

00. Huang Q, et al. A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. Nat Genet 46, 126-135 (2014).
01. Whitington T, et al. Gene regulatory mechanisms underpinning prostate cancer susceptibility. Nat Genet 48, 387-397 (2016).
02. Zhang P, et al. High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. Nat Commun 9, 2022 (2018).
03. Ahmed M, et al. CRISPRi screens reveal a DNA methylation-mediated 3D genome dependent causal mechanism in prostate cancer. Nat Commun 12, 1781 (2021).
04. Li K, Wei GH, Yin Y, Feng J. Targeting Noncoding cis-Regulatory Elements for Cancer Therapy in the Context of the 3D Genome. Cancer Discov 14, 2061-2065 (2024).
05. Schmitges, F.W. et al. Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res* **26**, 1742-1752 (2016).
06. Najafabadi, H.S. et al. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol* **33**, 555-62 (2015).
07. Kean, M.J. et al. Structure-function analysis of core STRIPAK Proteins: a signaling complex implicated in Golgi polarization. *J Biol Chem* **286**, 25065-75 (2011).
08. Seiler, C.Y. et al. DNASU plasmid and PSI:Biology-Materials repositories: resources to accelerate biological research. *Nucleic Acids Res* **42**, D1253-60 (2014).
09. Yu, F. et al. Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. *Mol Cell Proteomics* **19**, 1575-1585 (2020).
10. Teo, G. et al. SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *J Proteomics* **100**, 37-43 (2014).
11. Gao, P. et al. Biology and Clinical Implications of the 19q13 Aggressive Prostate Cancer Susceptibility Locus. *Cell* **174**, 576-589.e18 (2018).
12. Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-8 (2016).
13. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
14. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2021).
15. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
16. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).
17. Machanick, P. & Bailey, T.L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696-7 (2011).

18. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-8 (2011).
19. Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187-91 (2014).
20. Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T. & Flicek, P.R. The ensembl regulatory build. *Genome Biol* **16**, 56 (2015).
21. Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523 (2019).
22. Finn, R.D. et al. Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-30 (2014).
23. Castro-Mondragon, J.A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50**, D165-D173 (2022).
24. Bravo Gonzalez-Blas, C. et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* **20**, 1355-1367 (2023).
25. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401-4 (2012).
26. Bailey, T.L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res* **40**, e128 (2012).
27. Pickrell, J.K., Gaffney, D.J., Gilad, Y. & Pritchard, J.K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27**, 2144-6 (2011).
28. Degner, J.F. et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207-12 (2009).
29. Pickrell, J.K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-72 (2010).
30. Gong, J. et al. PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Res* **46**, D971-D976 (2018).
31. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
32. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977-D985 (2023).
33. Vivian, J. et al. Toil enables reproducible, open source, big biomedical data analyses. *Nat Biotechnol* **35**, 314-316 (2017).

Figures

Figure 1

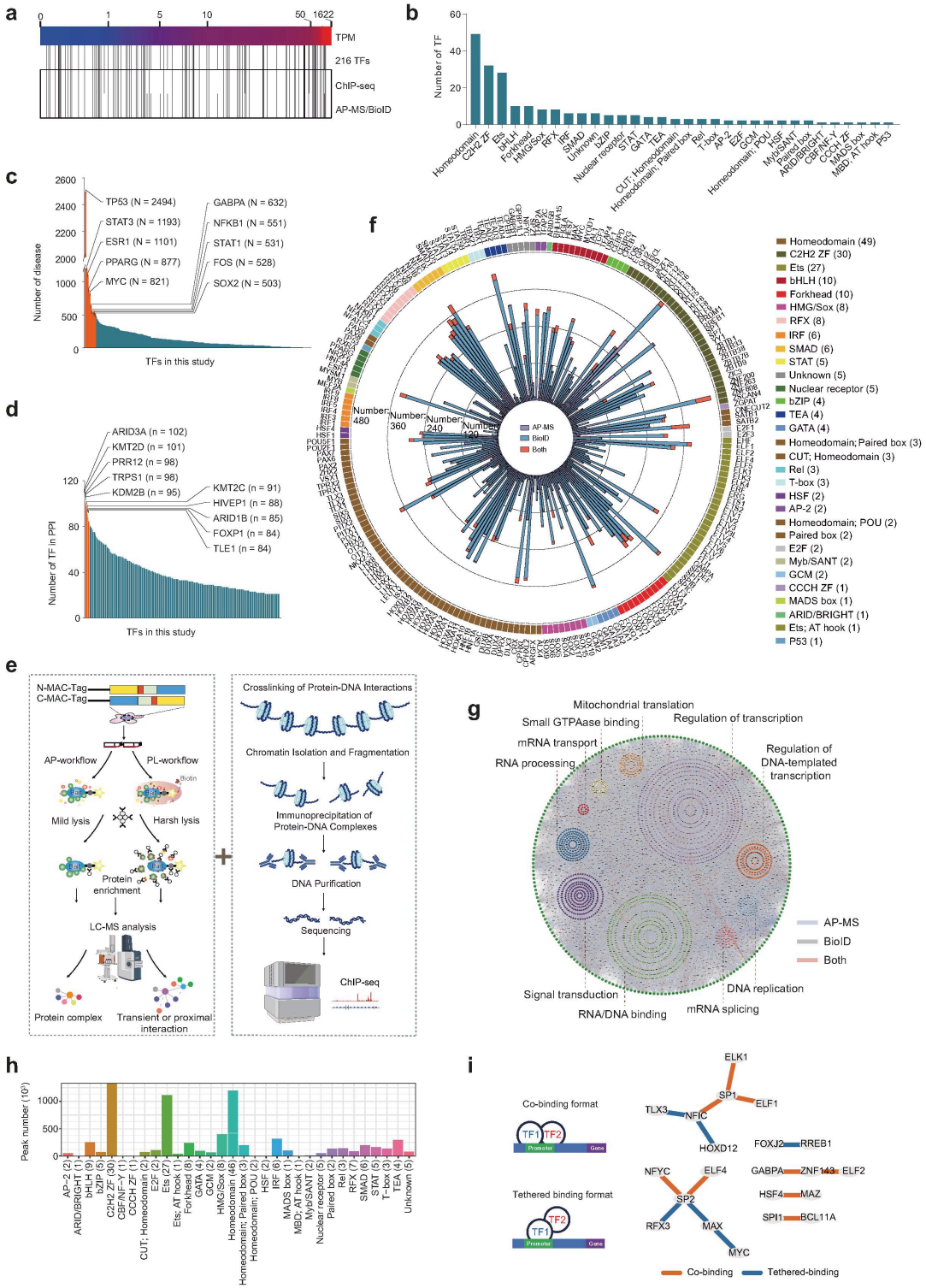


Figure 1

Proteocistromic landscape of transcription factors (TFs). (a) The 216 TFs profiled in HEK293 cells are stratified by expression level (TPM) and annotated according to the experimental modality applied, including ChIP-seq and AP-MS/BioID. **(b)** Distribution of TFs across DNA-binding domain (DBD) families. **(c)** Distribution of disease associations among TFs. TFs are ranked by the number of associated diseases in descending order,

with the top 10 highlighted by rectangular frames. **(d)** Distribution of TF-prey protein interactions. Prey genes are ranked by the number of interacting TFs in descending order, with the top 10 highlighted by rectangular frames. **(e)**Schematic overview of the integrated proteocistromic experimental workflow and downstream analyses. **(f)** Circos plot showing the distribution of protein-protein interaction (PPI) counts for individual TFs. **(g)** Functional enrichment analysis of TF interactors based on Gene Ontology (GO) biological processes and molecular functions, revealing broad involvement in regulatory and signaling pathways. All displayed GO terms are statistically significant ($P < 1 \times 10^{-3}$). To reduce redundancy, each protein was assigned to a primary GO term corresponding to its most representative function. **(h)** Total number of ChIP-seq binding peaks aggregated by DBD family. Numbers in parentheses indicate the number of TFs analyzed within each DBD family. **(i)** Evidence for TF complexes inferred from integrated proteomic PPI data and cistromic motif co-occurrenceanalysis. Co-binding interactions are shown in orange, and tethered-binding interactions in blue.

Figure 2

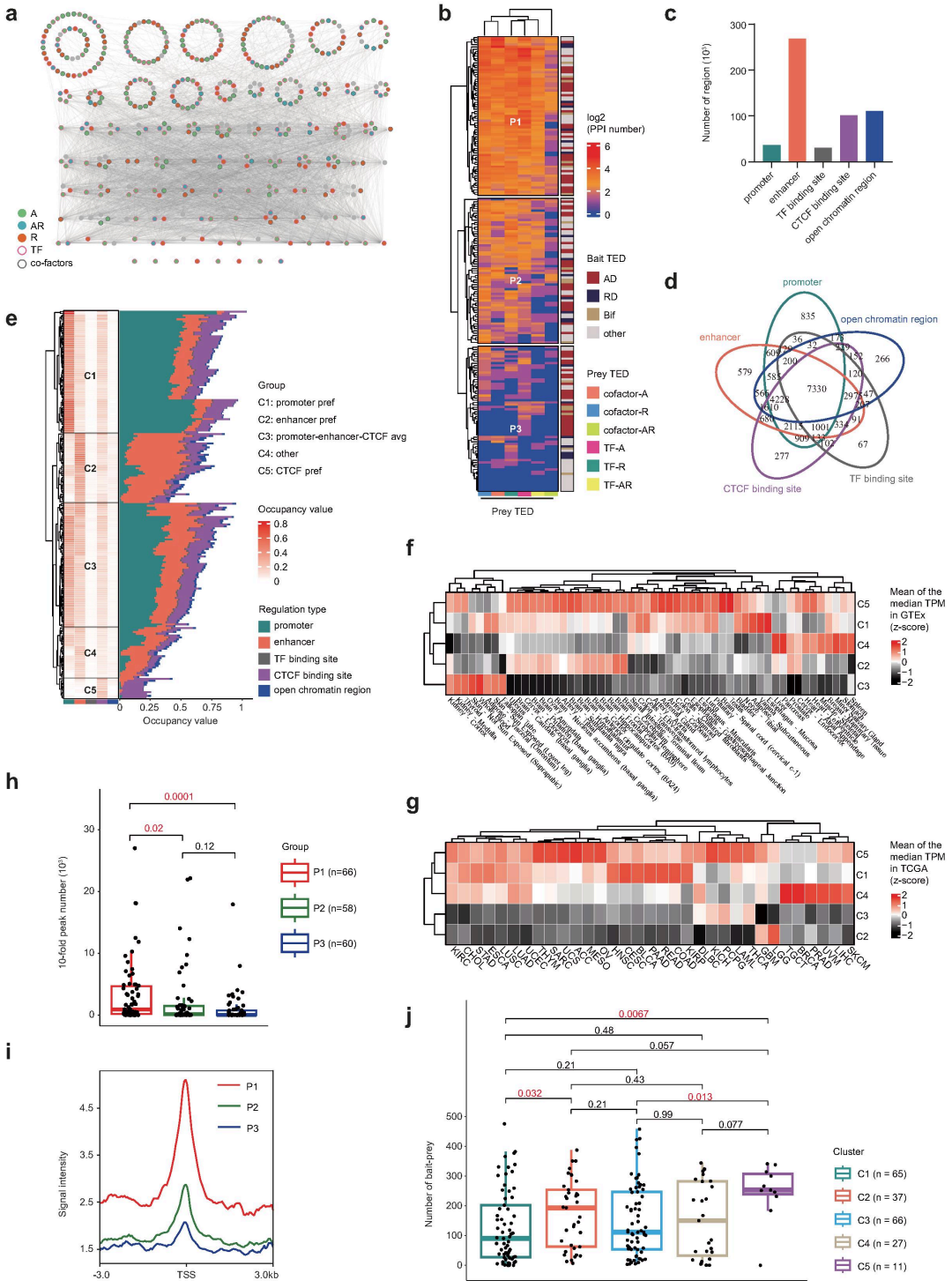


Figure 2

Integrated analysis of proteomic and cistromic coordination among transcription factors. **(a)** Network map of TF-TF and TF-cofactor interactions. Nodes represent individual TFs (pink) and cofactors (gray). Interactions are organized by force-directed layout clustering. Cofactors are annotated based on UniProt functional classification as activators (A, green), repressors (R, red), or dual-function regulators (AR, blue). **(b)** Clustering of TF effector

domain (ED) activity based on protein–protein interaction (PPI) profiles. TFs were grouped according to the number of PPIs with other TFs and cofactors, and classified as A, R, or AR based on UniProt annotations. Hierarchical clustering with a predefined number of clusters partitioned TFs into high-interaction (P1), medium-interaction (P2), and low-interaction (P3) groups. **(c)** Number of distinct genomic functional regions associated with TF binding sites. **(d)** Overlap of genes annotated to different classes of regulatory regions, showing intersections among gene sets assigned to promoter, enhancer, CTCF-associated, and other regulatory elements based on cistromic annotations. **(e)** Cistromic clustering of TFs based on the genomic distribution of their binding peaks across functional regions using hierarchical clustering, yielding five groups: promoter-preferred (C1), enhancer-preferred (C2), balanced across promoter–enhancer–CTCF (C3), other (C4), and CTCF-preferred (C5). **(f)** Comparison of TF expression across clusters C1–C5 in normal tissues using GTEx v8 data. For each tissue, median TF expression (TPM) was calculated across samples, and cluster-level expression was obtained by averaging median values of TFs within each cluster. The resulting matrix was z-score normalized and visualized as a heatmap. **(g)** Comparison of TF expression across clusters C1–C5 in tumor tissues using TCGA data, highlighting differential expression patterns among cistromic clusters. **(h)** Comparison of the number of high-confidence (≥10-fold enrichment) binding peaks among TFs in P1-P3. **(i)** Signal intensity of TF binding at transcription start site (TSS) regions for TFs in P1–P3. **(j)** Comparison of the number of PPIs associated with TFs across cistromic clusters C1-C5.

Figure 3

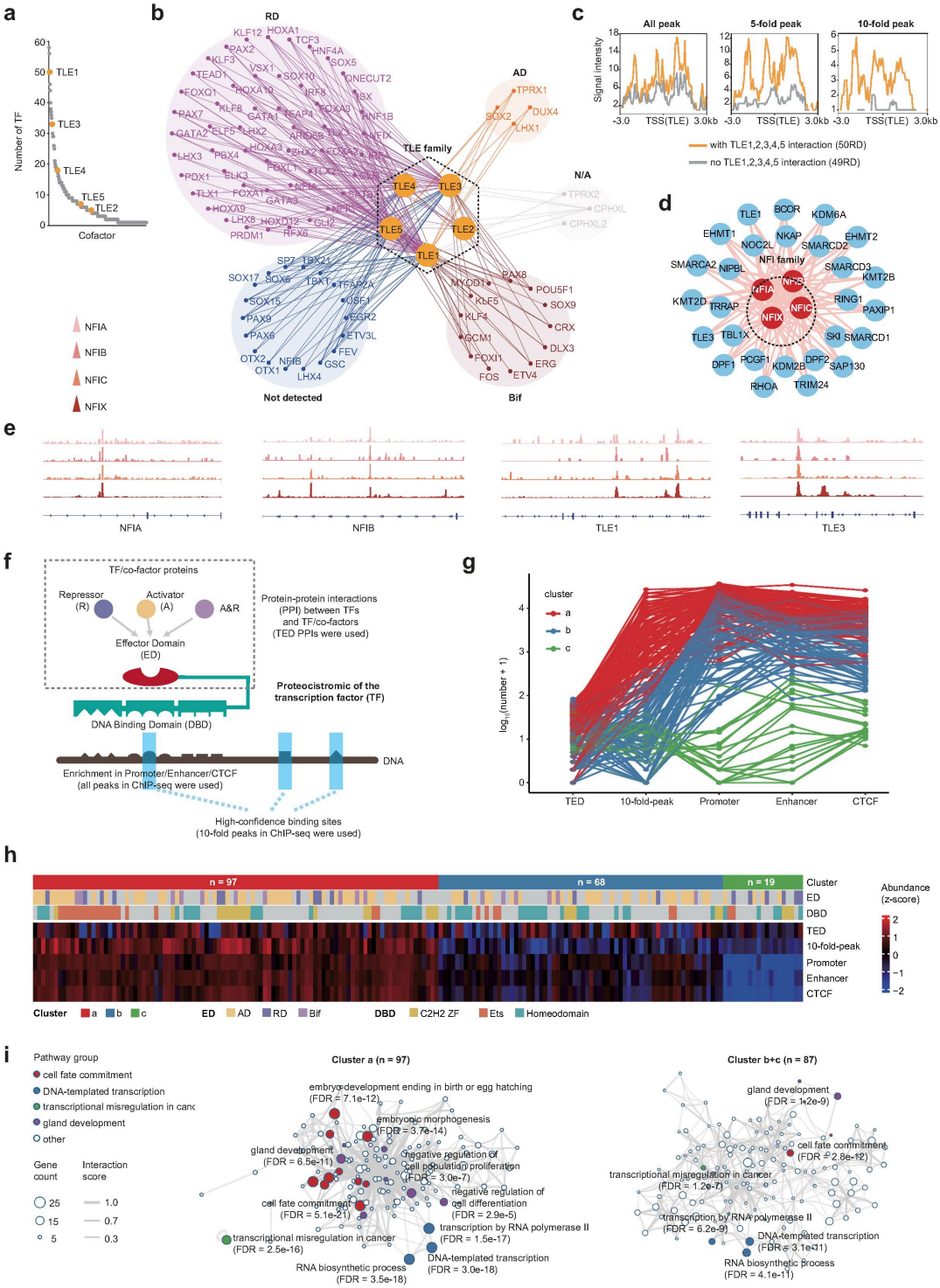


Figure 3

Classification of transcription factors based on integrated proteocistromic features. (a) Distribution of cofactors interacting with TFs derived from PPI data, with the TLE family highlighted in orange. **(b)** PPI network of TF-TLE family interactions. TFs are categorized by effector domain (ED) type: activation domain (AD), repression domain (RD), or bifunctional domain (Bif), which confers both activating and repressing activities; N/A

indicates unavailable annotation. **(c)** Signal intensity of RD-type TF binding at the TSS regions of TLE family target genes, stratified by TLE interaction status. Among the 99 RD TFs analyzed, orange traces represent 50 RD TFs that interact with TLE proteins, whereas gray traces represent 49 RD TFs without detected TLE interactions. **(d)** Interaction network of proteins associated with the nuclear factor I (NFI) family of TFs. **(e)** Genome-wide binding profiles of NFI family TFs at loci of associated genes exhibiting significant binding signals. **(f)** Schematic illustration of proteocistromic effect score calculation. The dashed box depicts quantification of TF effector domain (ED) effects, whereas the lower panel illustrates assessment of DNA-binding domain (DBD) effects. **(g)** K-means clustering of 184 TFs with quantified proteocistromic effect scores, partitioning TFs into three distinct categories. **(h)** Heatmap of proteocistromic features across the three TFs clusters. From top to bottom: number of TFs in each cluster; proteocistromic effect category (red, high effect; blue, intermediate effect; green, low effect); ED type annotation; DBD family annotation (highlighting the three most represented DBD families); and five quantitative indicators used to compute proteocistromic effects. All values are z-score normalized, with red indicating higher and blue indicating lower relative abundance. **(i)** Pathway enrichment network of TFs from different proteocistromic clusters. For cluster a and combined clusters b+c, the top 20 significantly enriched pathways (FDR < 0.05) are highlighted, with nodes sharing the same cluster ID positioned in close proximity.

Figure 4

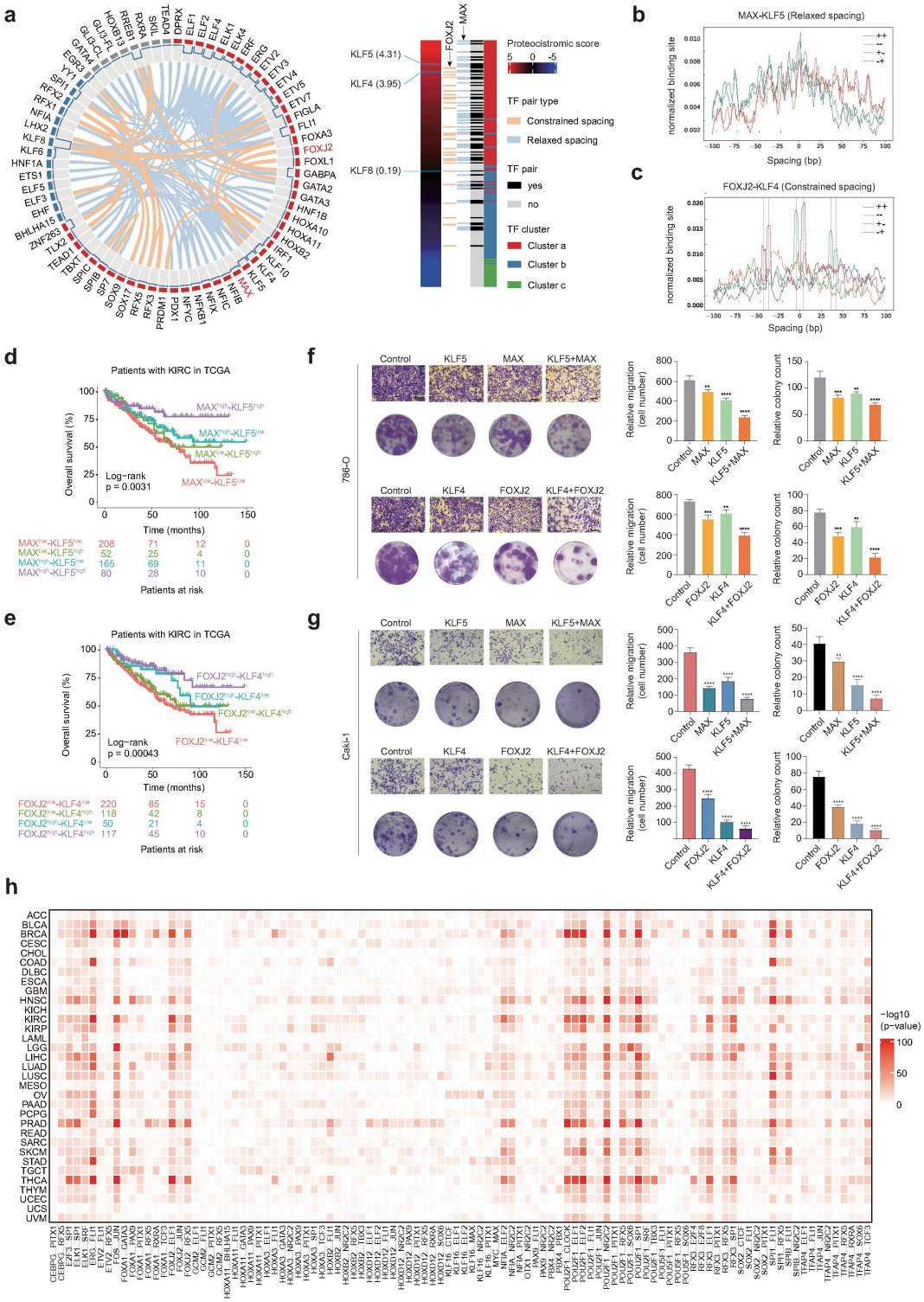


Figure 4

Functional characterization of transcription factorpairs. (a) Circos plots depicting relational networks of TFs capable of forming paired binding configurations. Colored squares on the outer ring indicate proteocistromic categories: red, cluster a (hot proteocistromic); blue, cluster b (cold proteocistromic); and gray, TFs without available proteocistromic annotations. Chords in the inner circle denote TF pair types, with light

blue indicating relaxed-spacing pairs and light yellow indicating constrained-spacing pairs. MAX and FOXJ2 are highlighted as the TFs with the highest proteocistromic scores among relaxed- and constrained-spacing pairs, respectively. **(b, c)** Spacing distributions of representative TF pairs, MAX-KLF5 **(b)** and FOXJ2-KLF4 **(c)**. "+" and "–" indicate motif orientation on the positive and negative DNA strands, respectively, and dashed lines denote significantly constrained spacing. **(d, e)** Kaplan-Meier survival analyses of kidney renal clear cell carcinoma (KIRC) patients stratified by combined expression levels of TF pairs MAX-KLF5 (**d**) and FOXJ2-KLF4 (**e**). Patients with expression above the cohort mean are classified as "high," and those below the mean as "low." **(f, g)** Representative images and quantification of cell migration and colony formation assays assessing cooperative effects of KLF4-FOXJ2 and MAX-KLF5 pairs in 786-O **(f)** and Caki-1 **(g)** cells. Scale bars, 100 μm. Data are shown as mean ± s.d. from n = 3 technical replicates. P values were calculated using two-tailed Student's *t* test; *** $P < 0.0001$. **(h)** Spearman correlation analysis of expression coherence for 104 TF pairs across multiple cancer types based on TPM-normalized RNA-seq data.

Figure 5

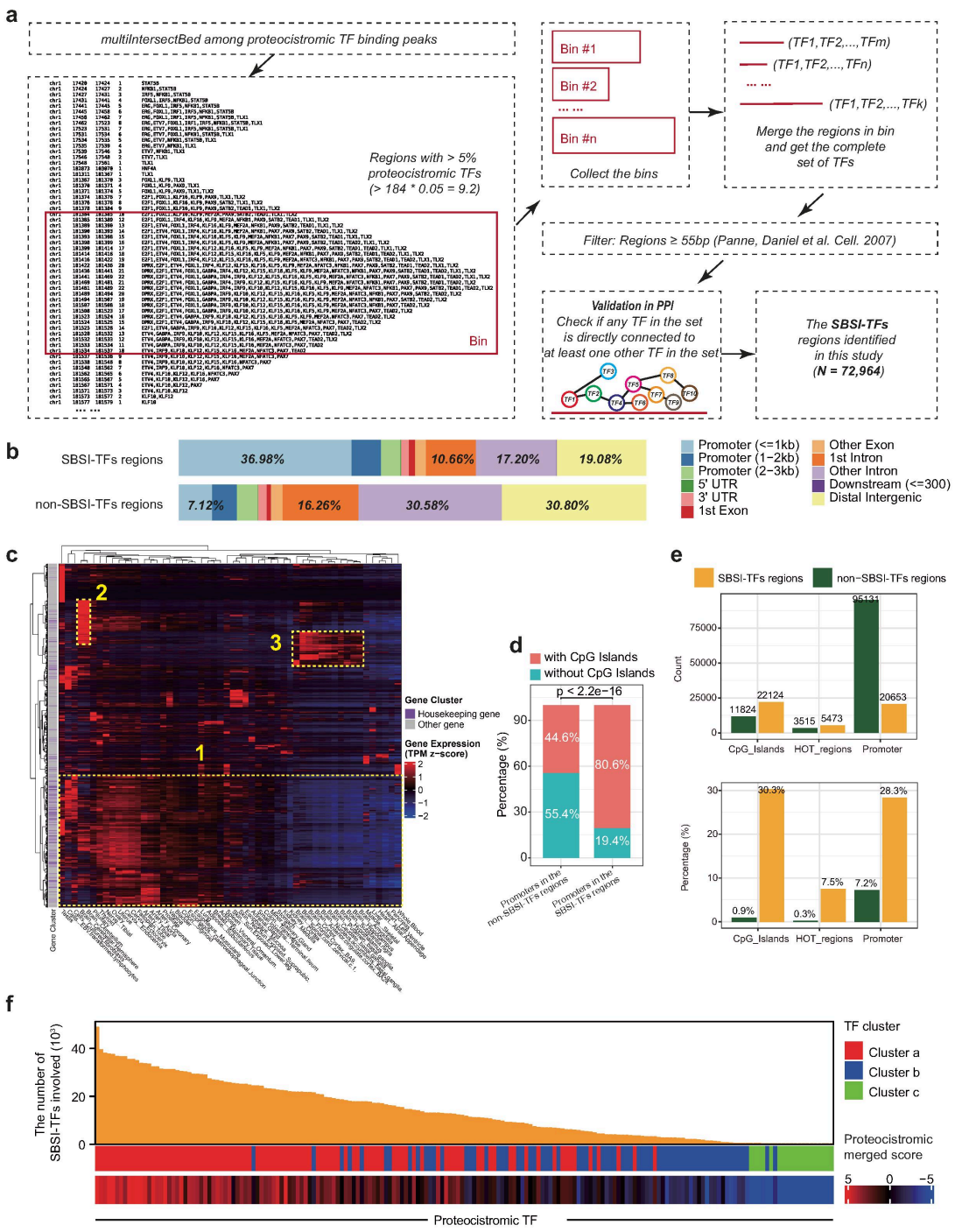


Figure 5

Identification and genomic characterization of SBSI-TFs derived from proteocistromic TFs. **(a)** Schematic of SBSI-TFs identification using integrated proteocistromic analysis. First, TF binding peaks from all proteocistromic TFs were merged using the *multiIntersectBed*function in BEDTools. Continuous genomic regions containing binding peaks from at least 5% of proteocistromic TFs (n = 10) were defined as bins. Bin lengths

were calculated, and bins shorter than 55 bp were excluded. Protein-protein interaction (PPI) data were then used to assess whether TFs co-occupying each bin were also physically interacting (i.e., all TF pairs within a bin exhibited PPIs). Bins meeting this criterion were defined as SBSI-TFs regions. **(b)**Genomic distribution of SBSI-TFs and non-SBSI-TFs regions across functional genomic elements. Non-SBSI-TFs regions were defined as proteocistromic TF binding bins identified by *multiIntersectBed* that did not satisfy the PPI validation criterion. **(c)** Heatmap of tissue-specific expression patterns of genes proximal to promoter-associated SBSI-TFs regions based on GTEx data. Housekeeping genes are highlighted in purple in the left annotation bar. Boxed regions indicate: Box 1, genes without strong tissue specificity (broad expression across tissues); Boxes 2 and 3, genes with significantly elevated expression in brain tissues. **(d)** Comparison of the proportion of promoters containing CpG islands between SBSI-TFs and non-SBSI-TFs regions. Statistical significance was assessed using Fisher's exact test. **(e)** Absolute counts (upper panel) and proportions (lower panel) of CpG islands, HOT regions, and promoters overlapping SBSI-TFs regions (yellow) and non–SBSI-TFs regions (green). **(f)** Quantification of SBSI-TFs regions involvement for each proteocistromic TF. TFs are ranked by the number of SBSI-TFs regions in which they participate. The lower panel indicates corresponding cluster assignments and merged proteocistromic effect scores for each TF.

Figure 6

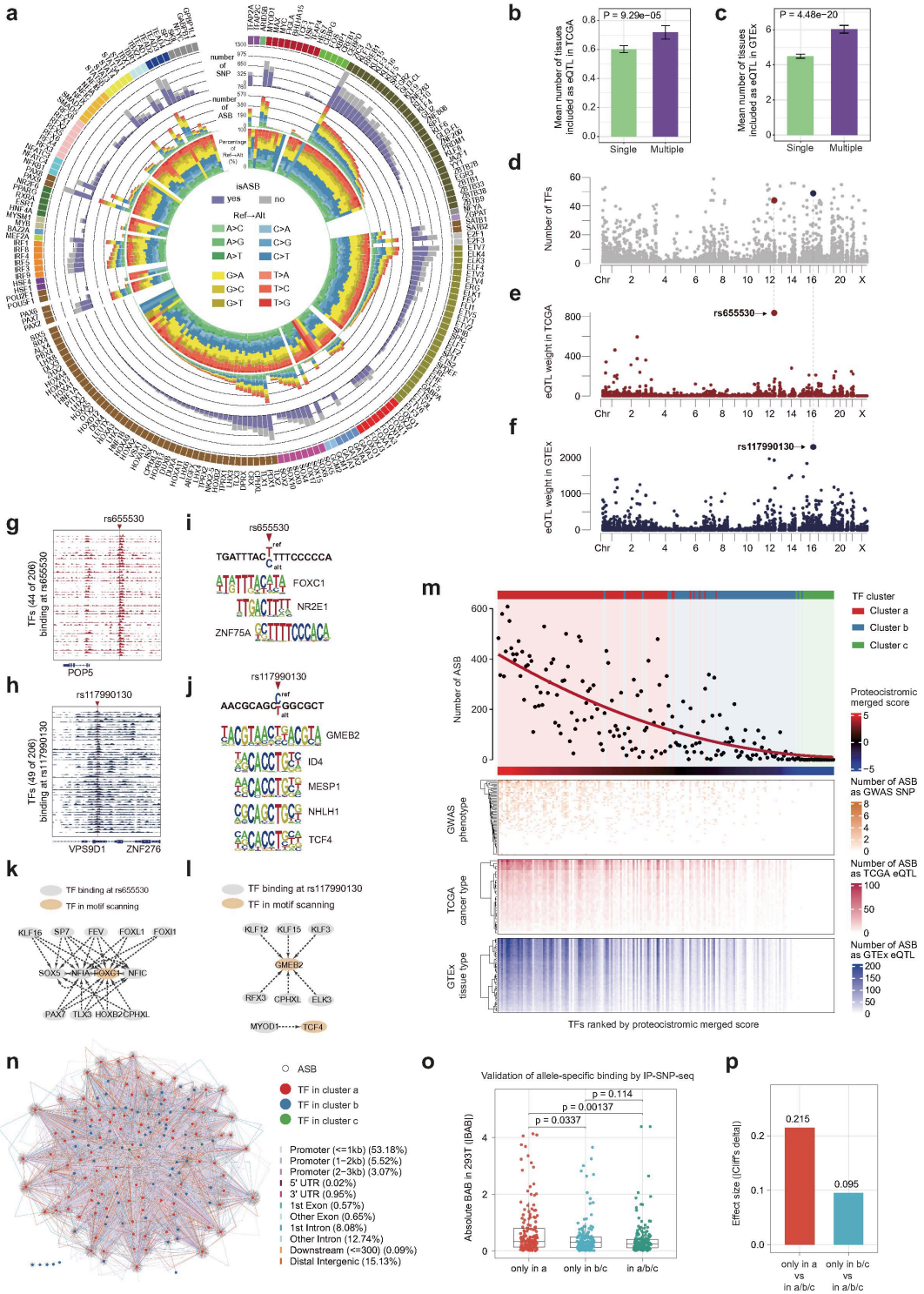


Figure 6

Analysis of allele-specific binding (ASB) sites associated with proteocistromic TFs. (a) Circos overview of ASB features across 206 TFs. From outer to inner rings: TFs grouped by DNA-binding domain (DBD) family; counts of ASB and non-ASB events per TF; absolute numbers of nucleotide variants exhibiting ASB for each TF; and percentages of nucleotide variants showing ASB per TF. **(b, c)** Comparison of the number of tissues in which ASBs

act as expression quantitative trait loci (eQTLs) when bound by a single TF versus multiple TFs. “Single” denotes ASBs associated with only one TF, whereas “Multiple” denotes ASBs associated with two or more TFs. (**b**) eQTLs from TCGA tumor tissues. (**c**) eQTLs from GTEx normal tissues. **(d)** Genome-wide distribution of ASBs identified in this study, where the y-axis indicates the number of TFs associated with each ASB.**(e, f)** Weight distribution of ASBs as eQTLs across human tissues, defined as: eQTL weight = (number of TFs associated with an ASB) × (number of tissues in which the ASB functions as an eQTL). ASBs with the highest weights are highlighted for TCGA **(e)** and GTEx **(f)**. **(g, h)** Genome browser views of TF binding enrichment at the highest-weight eQTLs identified in **(e)** and **(f)**. rs655530 for TCGA **(g)** and rs117990130 for GTEx **(h)**. **(I, j)** Motif analysis at ASB loci using the EEL algorithm for rs655530 **(i)** and rs117990130 **(j)**. **(k, l)** Protein interaction networks of TFs whose motifs are detected at ASB loci based on ChIP-seq enrichment for rs655530 (**k**) and rs117990130 (**l**). **(m)** Upper panel: scatter plot showing the number of ASBs associated with each proteocistromic TF, ordered by decreasing proteocistromic score, with a linear regression fit shown in red. Lower panel: heatmap displaying ASB counts per TF that are annotated as GWAS SNPs (orange), TCGA eQTLs (red), and GTEx eQTLs (blue), clustered by phenotype, cancer type, and tissue type. **(n)** TF-ASB network illustrating relationships between proteocistromic TFs and ASB loci. ASBs are shown as hollow black nodes, whereas TFs from clusters a, b, and c are shown as solid red, blue, and green nodes, respectively; edges indicate genomic associations between TFs and ASBs. **(o)** Validation of allele-specific binding by IP-SNP-seq in HEK293 cells. Boxplots with overlaid points show absolute binding allelic bias (|BAB|) for ASB SNPs stratified by motif cluster membership: only in cluster a, only in clusters b/c, and in both cluster a and clusters b/c. Each dot represents one SNP. Boxes indicate median and interquartile range (IQR); whiskers extend to 1.5× IQR. P values are from two-sided Wilcoxon rank-sum tests with Benjamini-Hochberg correction across three pairwise comparisons and are shown above brackets. **(p)** Effect sizes of allelic imbalance differences between motif-defined ASB subsets. Bars represent absolute Cliff’s delta values comparing “only in a” versus “in a/b/c” and “only in b/c” versus “in a/b/c”, respectively. Numeric labels indicate estimated effect sizes, with larger values reflecting greater separation of |BAB| distributions.

Figure 7

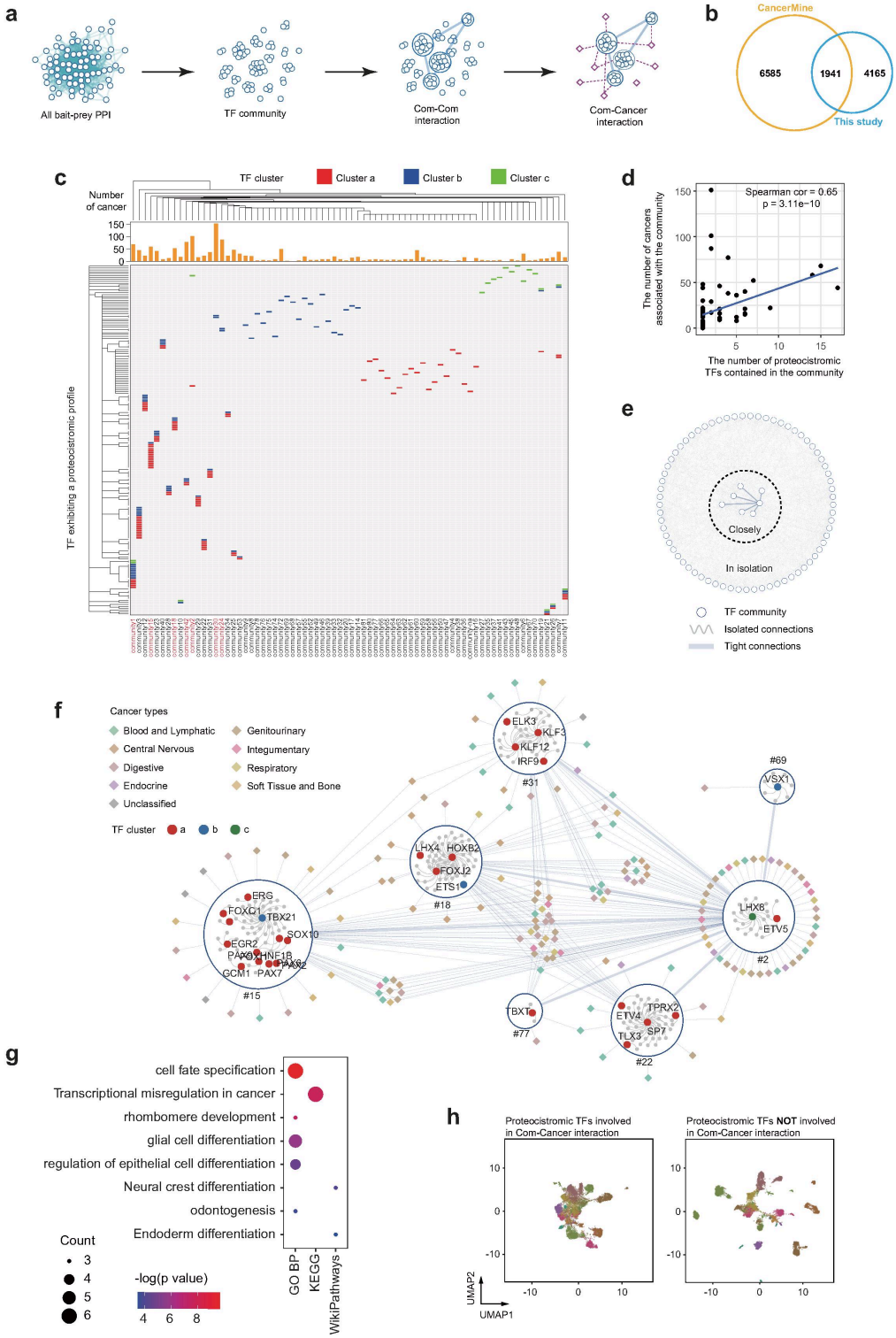


Figure 7

Associations between TF protein communities and cancer types. (a) Schematic overview of the workflow for identifying community-cancer associations. PPI data were clustered into protein communities using the Markov Cluster (MCL) algorithm, followed by assessment of inter-community connectivity. Significant community-community interactions were identified using Fisher's exact test with false discovery rate (FDR) < 0.05, focusing on

closely linked community pairs. Proteins within each community were then subjected to co-enrichment analysis with cancer-associated genes curated in the CancerMine database using the hypergeometric test (FDR < 0.05), thereby defining community–cancer associations. **(b)** Overlap between genes in the TF-centered PPI network and cancer-associated genes curated in CancerMine. **(c)** Heatmap of associations between communities containing proteocistromic TFs and cancer types. Rows represent proteocistromic TFs and columns represent communities. The bar plot above the heatmap indicates the number of cancer types associated with each community. Hierarchical clustering was performed using the complete-linkage method. **(d)** Spearman correlation between communities containing proteocistromic TFs and the number of associated cancer types. **(e)** Network of significant community-community interactions. Communities are shown as blue circles (n = 76). Straight edges indicate closely connected communities, whereas wavy edges denote mutually exclusive relationships. The central cluster highlights a collection of highly interconnected communities. **(f)** Network representation of tightly connected community modules associated with cancer. Diamonds of different colors indicate cancer types, and proteocistromic TFs within each community are highlighted (n = 28). **(g)** Pathway enrichment analysis of proteocistromic TFs from panel (c), showing significantly enriched pathways (FDR < 0.05). **(h)** UMAP projection of expression profiles of proteocistromic TFs across TCGA patient cohorts. The left panel uses TFs from the cancer-associated community collection shown in (**f**), whereas the right panel uses proteocistromic TFs not included in (**f**). Cancer types are grouped by organ systems as follows: Blood and lymphatic (DLBC, LAML, THYM); central nervous system (GBM, LGG, PCPG); digestive (CHOL, COAD, ESCA, LIHC, PAAD, READ, STAD); endocrine (ACC, THCA); integumentary (HNSC, SKCM, UVM); reproductive (BRCA, CESC, OV, PRAD, TGCT, UCS); respiratory (LUAD, LUSC); soft tissue and bone (SARC); urinary (BLCA, KICH, KIRC, KIRP, UCEC); and unclassified (MESO).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS3.xlsx
- TableS7.xlsx
- TableS1.xlsx
- TableS2.xlsx
- TableS6.xlsx
- TableS5.xlsx
- TableS4.xlsx
- WangzxetalFigureS1S9.pdf

- SupplementaryFigureLegends.docx